DOCUMENT RESUME

ED 181 080 TM 010 041

TITLE Testing, Teaching and Learning. Report of a Conference on Research on Testing (August 17-26, 1979).

INSTITUTION National Inst. of Education (DHEW), Washington, D.C.

PUB DATE Oct 79

NOTE 433p.: Appendix contains extracts from The National Conference on Achievement Testing and Basic Skills, March 1-3, 1978 (printed February 1979)

EDRS PRICE MF01/PC18 Plus Postage.

DESCRIPTORS *Achievement Tests; *Basic Skills; Cognitive Processes; Culture Free Tests; Dictionaries; Educational Diagnosis; Educational Objectives; *Educational Research; Educational Trends; Electronic Data Processing; Language Arts; Learning Processes; Mathematical Concepts; Reading Achievement; *Student Testing; Teaching Methods; Technological Advancement; *Testing Problems; Writing Skills

IDENTIFIERS Pennsylvania (Phildelphia); Philadelphia School District PA; Project TORQUE

ABSTRACT

The proceedings of the Conference on Research on Testing are presented, including the chairmen's and committees' reports. Invited papers are entitled: Educational Objectives and Educational Testing; Technology; The Sciences and Technologies; Mathematics; Mathematics: A View from the Schools; Information Processing Analyses of Mathematical Problem Solving; Reading; Writing: A School Perspective on Language; Automated Dictionaries; Some Emerging Trends in Testing; Cultural Consideration: African-American; Cultural Considerations: Hispanic-American; Instruction and Testing in the Philadelphia Public Schools; Project TORQUE: Diagnostic Models in Basic Mathematical Skills; and Effectiveness Measure in Reading. Extracts of some of the presentations at the National Conference on Achievement Testing and Basic Skills are appended. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED181080

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

RECEIVED
ERIC/TM

TESTING, TEACHING AND LEARNING

Report of a Conference on
Research on Testing

TM010 041

The National
Institute of
Education
U.S. Department of
Health, Education and Welfare
Washington, D.C. 20208

CONFERENCE PARTICIPANTS

GREGORY R. ANRIG, Commissioner of Education, State of Massachusetts
CARL BEREITER, Dept. of Applied Psychology, Ontario Institute for Studies in Education
JOHN SEELY BROWN, Xerox Palo Alto Research Laboratories
SYLVIA CHARP, Instructional Systems Divn., School District of Philadelphia
ALLAN COLLINS, Center for the Study of Reading, Bolt Beranek & Newman
ANN COOK, Community Resources Institute, New York, N.Y.
ALONZO A. CRIM, Superintendent, Atlanta Public Schools
PARKER DAMON, Principal, McCarthy-Towne School, Acton, Mass.
ROBERT B. DAVIS, Curriculum Laboratory, University of Illinois, Urbana
LAWRENCE T. FRASE, Bell Laboratories, Piscataway, N.J.
NORMAN O. FREDERIKSEN, Division of Psychological Studies, Educational Testing Service
ANDREW M. GLEASON, Dept. of Mathematics, Harvard University
DONALD ROSS GREEN, Director of Research, CTB/McGraw Hill, Monterey, Calif.
JAMES G. GREENO, Learning R&D Center, University of Pittsburgh
BARBARA HAYES-ROTH, Rand Corporation, Santa Monica, Calif.
LEON A. HENKIN, Dept. of Mathematics, University of California, Berkeley
ASA G. HILLIARD, School of Education, San Francisco State University
C. DALTON JONES, Dept. of Psychology, University of Massachusetts, Amherst
DAVID KLAHR, Dept. of Psychology, Carnegie-Mellon University
BERTRAM L. KOSLIN, Touchstone Applied Science Associates, Elmsford, N.Y.
GEORGE A. MILLER, Dept. of Psychology, Princeton University
AMADO M. PADILLA, Dept. of Psychology, Univ. of California, Los Angeles
ANDREW C. PORTER, Institute for Research on Teaching, Michigan State Univ.
FRANCES QUINTO, National Education Association, Washington, D.C.
MARIA RAMIREZ, Asst. Commissioner, N.Y. State Education Department
JUDAH L. SCHWARTZ, Div. for Study and Research in Education, Massachusetts Institute of Technology
SUSAN S. STODOLSKY, Depts. of Education and Human Development, University of Chicago
ROSS TAYLOR, Mathematics Consultant K-12, Minneapolis Public Schools
RALPH W. TYLER, Senior Consultant, Science Research Associates
SHELDON H. WHITE, Dept. of Psychology & Social Relations, Harvard University
JERROLD R. ZACHARIAS, Education Development Center

TESTING, TEACHING AND LEARNING

Report of a Conference on Research on Testing

August 17-26, 1979

Ralph W. Tyler
and
Sheldon H. White
Chairmen

U.S. Department of Health, Education, and Welfare
Patricia Roberts Harris, Secretary
Mary F. Berry, Assistant Secretary for Education

National Institute of Education
Michael Timpane, Acting Director

October 1979

Points of view or opinions expressed in this report are not necessarily those of the National Institute of Education or the Department of Health, Education, and Welfare.

FOREWORD

The National Institute of Education (NIE), the primary federal agency for educational research and development, sponsored the conference reported on here with the aim of fostering research into better means of assessing what students have learned and the difficulties they are having in learning. Such assessments are central to improving the practice of education and promoting educational equity - two of the Institute's main goals. The conference, held in August 1978, was a response to the widespread interests in research on testing--interest made clear by participants in a more broadly representative Conference on Achievement Testing and Basic Skills, convened by the Secretary of Health, Education, and Welfare earlier in 1978.

The discussion of issues and opportunities in testing and the research agendas in this report are proving very useful to NIE in developing a research agenda in testing as part of its Program on Teaching and Learning. We believe they will also be interesting and useful to a variety of public and private organizations and individuals concerned with improving education, with testing, and with the development of tests.

Michael Timpane, Acting Director
National Institute of Education

CHAIRMEN'S PREFACE

Until the late 1960s, criticism of educational testing was largely confined to academic debate. Since then, testing has become a center of public controversy.

Minority groups object that published tests are unfair to their children. Many teachers, parents, and school administrators find existing tests unsatisfactory as indicators of a student's educational progress. The National Education Association has called for a moratorium on testing.

In March 1978, Secretary of Health, Education, and Welfare, Joseph A. Califano, Jr., convened the three-day National Conference on Achievement Testing and Basic Skills. (A report on the conference will be found in the Appendix of this volume.) Participants were chosen to be broadly representative of current concerns. One goal of the conference was to find ways in which HEW could help states and localities to use tests more effectively. Another was to open a national discussion about the reasonable use of tests to improve the quality of elementary and secondary education. Among its recommendations, the conference called for expansion in scope of Federal support of research on learning and on the uses of testing.

In August 1978, responding to this recommendation, the National Institute of Education, HEW's educational research agency, sponsored a ten-day conference on research on testing, planned by four members of the NIE staff: Sylvia Scribner, then Associate Director for Teaching and Learning, John M. Mays and Arthur S. Melmed in the Office of the Director, and Jeffry Schiller, Assistant Director for Testing, Assessment, and Evaluation.

The thirty-one participants at the conference comprised persons concerned with teaching and educational administration and policy, with research and development relevant to education, with educational testing, with various areas of educational content, and with information-handling technology.

Many of the points made at the first conference were also strongly asserted at the second conference:

- A national achievement test is no more the answer to improving educational quality than is a national curriculum.
- Beware of tests that are hastily constructed to meet demands that are hastily made.
- Everyone needs better information on what tests can and cannot do - test users, test takers (including their parents), and test makers.
- Current testing procedures are unfair to particular minority groups.
- Current testing procedures are not helpful to teachers or students in their day-to-day efforts to teach and learn.

- Tests are not the whole story in assessing educational progress and ways must be found to make better uses of qualitative information from other sources.

How, then, is research to help improve our use of tests? The general strategy of the conference was to address both short-term realities and long-term possibilities. As might be expected, the scope of the discussion was broad, ranging from present-day problems to test usage in schools to the state-of-the-art of computerized testing and teaching. Particular emphasis was placed on exploration of the implications for testing of three new areas of progress: First, recent advances in cognitive science which provide new understanding of human intellectual processes and allow better design of instruction and testing and better fitting of them to cultural background. Second, progress in information-handling technology, which is in a period of revolutionary increase in capability and lowering cost, allowing greatly increased breadth and individualization of instruction and testing. Third, the new sense of what is achievable in education that has grown out of collaboration between teachers and scholars in the last two decades.

The conference began with presentations of prepared papers, which are reproduced in Part Two of this volume. The conference then divided into three working groups, each reflecting the same diversity of interests and backgrounds as the conference as a whole. The working groups met separately, reporting to each other periodically at plenary sessions, and they wrote reports which make up Part Three of this volume.

The Chairmen's Report, Part One of this volume, which was prepared with the help of Joseph Turner, sets forth from our perspective the deliberations and recommendations of the conference. All participants have reviewed this document and their comments are incorporated in it or are appended.

Ralph W. Tyler
Director, Emeritus
Center for Advanced Study in
the Behavioral Sciences

Sheldon H. White
Dept. of Psychology
Harvard University

CONTENTS

Foreword iii

Chairmen's Preface v

PART ONE: CHAIRMEN'S REPORT 1
  Summary 3
  I. The Issues Concerning Tests 7
    A. Major Uses of Current Tests 7
    B. Major Criticisms of Current Tests 8
  II. Recommendations 11
    A. Bettering Fitting of Testing to Cultural Background of Students 11
    B. Better Fitting of Testing to Educational Objectives 12
    C. Combining Testing with Teaching 22
  References 28
  Individual Comments 29

PART TWO: INVITED PAPERS 33
  I. Introduction 35
  II. Educational Objectives and Cognitive Models 53
  III. Testing 185

PART THREE: COMMITTEE REPORTS 335
  Testing to Improve Teaching and Learning: The View from Cognitive Science 337
  Teaching, Testing, and Learning: The Approach Through Research on Classroom Processes 357
  Hard Decisions from Soft Data 375

APPENDIX: THE NATIONAL CONFERENCE ON ACHIEVEMENT TESTING AND BASIC SKILLS 391

PART ONE: CHAIRMEN'S REPORT

SUMMARY

The Issues Concerning Tests

Broadly speaking, educational tests are used in American education today to achieve four purposes:

1. Holding teachers, schools, and school systems accountable.
2. Making decisions concerning individual students.
3. Evaluating educational innovations and experimental projects.
4. Providing guidance to teachers in the classroom.

The major current criticisms of educational tests are the following:

1. Tests do not reflect the full range of student cultural backgrounds and thus lead to decisions that are unfair to minority students.
2. Tests have only limited value for holding educators accountable.
3. Tests exercise a limiting effect on classroom teaching.
4. Tests are too narrow in scope to provide for fair evaluation of new approaches to teaching.

Recommendations

The conference participants believe that testing, and thereby education, can be substantially improved through a variety of research and development and other related activities responding to the issues above. Recommendation A responds to the first criticism. Recommendation B responds to the last three criticisms, which arise in considerable part from a lack of fit between testing and educational objectives Recommendation C looks toward a system in the future in which testing is made more effective and helpful by being merged into the teaching process.

A. Better Fitting of Testing to Cultural Background of Students

Development of equivalent culture-specific tests in which the difficulty level and knowledge domains are the same, but the language and illustrative material are tailored to the background knowledge of particular groups.

B. Better Fitting of Testing to Educational Objectives

1. Criterion-referenced Testing

Increasing research and development on the theory and practice of criterion-referenced tests, which are specifically designed to determine what a student knows and can do in a specified domain of educational objectives.

2. Information-handling Technology

Application to testing of the increasingly available low-cost information-handling technology in such areas as:

a. Computer-based Item Pools

Creation of computer-based pools of test items by educational centers and test publishers which would allow teachers, schools, and school systems to create tests closely matched to their educational objectives.

b. Testing Problem-solving Processes

Development of means of testing in one of the principal aims of education - the ability to address more complex problems of the sort encountered in work and personal like - now made possible by new low-cost computer capability.

c. Tailored Testing

Development of more tests in which the computer matches the level of difficulty of items to the capability demonstrated by students and thus provides better and quicker assessments.

3. Providing Better Information on Tests and Testing

Developing and making available the information on tests and testing that educators, parents, policymakers, and the public need to know in fitting testing to educational objectives:

a. Clarifying Basic Conceptions of Testing

Taking a new look at what we can and cannot expect from tests in the light of present understanding of the complex nature of human knowledge - and making the results easily available to interested parties.

b. Appropriate Use of Tests in Education

A study of the use of tests in American education documenting the typical uses of tests and investigating the effects of testing on teaching, on the education of individuals, on educational innovation, and on budgetary and other decisions - to provide much firmer information on the suitability of various testing practices than is now possible.

c. Information on Published Tests

Formation of privately supported consumer groups as sources of information about the availability and capabilities of various tests, and provision for fully informative disclosure of test items after administration.

C. Combining Testing with Teaching

Sustained research and development over the next decade or more toward a system in which testing is merged into the teaching process and provides timely and rich feedback to the student, the teacher, and other appropriate parties. Four elements are seen as essential to this R&D program:

1. Cognitive Science

Cognitive science, an interdisciplinary movement comprising psychologists, computer scientists, philosophers, linguists, anthropologists, and educators, is providing powerful new understanding of the processes in learning and performing the intellectual tasks to be tested.

2. Interactive Teaching-Testing and Technology

The rapidly increasing capacity and decreasing cost of electronic information-handling technology make feasible branching, exploratory interplay with the learner, of which testing can be a natural part.

3. Subject Matter

Involvement of knowledgeable teachers and scholars in test making, important for any test, is essential in the development of teaching-testing systems and tests of problem-solving processes.

4. The Natural Classroom Situation

The teaching-testing systems envisioned are intended as aids to teachers rather than substitutes for them and thus must be developed making full use of what is known about the way instruction and instructional guidance take place at the grade level and in the subject area in question and about the successes and failures of previous innovations.

I. THE ISSUES CONCERNING TESTS

The theory that has, explicity or implicitly, dominated most thinking about tests is that they provide measures, analogous to length and weight, that can be used to arrange children and what they have learned, in a simple linear order. This theory supported initial research on testing, but it is obviously inadequate to our present conceptions of the complex nature of human knowledge.

George A. Miller

The widespread use of tests suggests that they fill the needs of many people. Who uses tests? When? For what purposes? Although there is now much criticism of testing, a careful observer will note that many critics examine specific issues rather than find fault across the board and that different critics pose different objections. We consider first the major uses of current tests and then the chief criticisms leveled at them.

A. Major Uses of Current Tests

Broadly speaking, educational tests are used in American education today to achieve four purposes: accountability, selection, evaluation of experimental projects, and instructional guidance.

1. Tests are used to hold teachers, schools, and school systems accountable.

Many principals, superintendents, and other educational authorities use test scores, particularly scores on achievement tests, as a rough gauge of the adequacy of the performance of a teacher, a school, or larger administrative unit. Parents, voters, and legislators also use such information in judging schools and school systems. The results of a test are taken to indicate the amount of learning accomplished by the average student in a classroom or larger unit. Educational authorities may then decide that the teaching offered is either adequate, inferior, or superior.

2. Tests are used to make decisions concerning individual students.

Educational authorities use the same tests when placing individual students in special programs and classes and in counselling them on plans for future education and careers. Procedures vary from one system to another, but combinations of ability and achievement tests are generally used for these purposes. Authorities may decide to assign a pupil in the early grades to a slow paced program if both ability and achievement scores are low, or to a program for the gifted if both are high. A low score in reading achievement coupled with a high ability score raises the possibility of a reading or learning disability; more specialized diagnostic tests may then be used.

College admissions officers depend heavily on student scores on the Scholastic Aptitute Test (SAT) and similar tests in considering applicants. Some colleges make the achievement of a score above some cutoff point a necessary but not sufficient condition for admission. School grades, letters of recommendation, and interviews are significant, but exactly which factors are considered and how they are weighed varies from college to college and from time to time.

3. Tests are used to evaluate educational innovations and experimental projects.

Government agencies, private foundations, and school systems sponsor experimental projects in American schools and seek to evaluate these projects through use of standardized achievement tests. A recent wave of experimental projects was the curriculum reform movement in science and mathematics, which began in the 1950s. Another larger wave came in the 1960s when widespread efforts were made to improve the education of children from backgrounds of poverty and discrimination. Evaluators of experimental projects continue to wrestle with the task of matching tests to project objectives. In some cases, experimenters have found available tests unsuited to their projects and have developed new ones.

4. Tests are used to provide guidance to teachers in the classroom.

Test makers and the educators who select published tests for use in schools exert influence directly and indirectly on teachers in the classroom. Direct influence is exerted by the choice of subtests to measure the strengths and weaknesses of students in particular component skills. Ideally, a teacher studying the scores could determine where to invest teaching time with particular students or classes. The indirect influence results when tests are used for accountability. If teachers are to be rated by administrators on the basis of their students' scores, prudent teachers will emphasize in their teaching the topics emphasized in the tests.

B. Major Criticisms of Current Tests

The major present-day criticisms of educational testing are the following:

1. Tests do not reflect the full range of student cultural backgrounds and thus lead to decisions that are unfair to minority students.

The underlying logic of standardized testing requires that a given test performance must have the same meaning for all children or groups of children being assessed or compared. We would not have much confidence in a thermometer if its operation varied with certain background characteristics of patients that were irrelevant to their physical condition. Similarly, our confidence that educational assessments are accurate and selection decisions fair rests on the condition that test performance is not affected in any systematic way by irrelevant background characteristics of children - whether they are boys or girls, black or white, or of Hispanic, Italian or Irish origin. Yet there is increasing consensus among psychologists, educators and some test makers, as well as members of minority groups, that this essential condition of "sameness" is not met by the ability and achievement tests most widely used in education. These tests commonly reflect the vocabulary specializations, language styles, and cultural knowledge of English-speaking children of the majority culture. When such tests are used with children from other cultural and linguistic groups, these children are at a disadvantage. On the most simple level, we can understand that a test of history knowledge or reading skills written in English places one set of demands on children who are native English speakers and quite another set of demands on the estimated 2.5 million school-aged children in this country with limited English proficiency. But even for children who are proficient in English, discrepancies between their everyday home language and test language or between their common knowledge and the knowledge presupposed by the test are sources of unknown and uncontrolled variability in the testing situation.

Other factors varying with cultural background - motivation to perform, familiarity with test-taking, responses to stress - are known to affect performance in test-like experimental situations. In a culturally pluralistic society such as the United States, testing practices that reflect the dominant culture's standards of language function and shared knowledge and behavior provide imperfect measurement instruments for other groups. But existing tests are publicly interpreted as though they are culture-fair, and low test scores are used to stigmatize minority children, to justify negative interpretations of their abilities and negative decisions about their schooling.

2. Current standardized tests have only limited value for holding teachers, schools, and school systems accountable for the quality of education.

The use of current standardized test to evaluate the effectiveness of education is under attack by educators and other persons concerned with education. They argue that the educational objectives tested often differ from what the school is seeking to teach. Further, it is pointed out that the tests have generally been designed to sort students for various administrative purposes rather than to determine how well they have learned what is being taught. Similar criticism is implied by the call in many States for competency standards that are enforced through new tests of performance in carrying out tasks drawn from ordinary life.

3. Tests exercise a limiting effect on classroom teaching.

Several national educational groups have called for a moratorium on testing. It is argued that standardized tests have no positive direct usefulness in guiding instruction, and their indirect influence - implicitly laying down goals and standards - disrupts or blocks teaching. Despite inclusion in the published tests of various subtests to identify a student's strengths and weaknesses, critics say the categories are so broadly defined, the tests are given so infrequently, and the time from test administration to report of results to teachers is so long that tests do not help teachers in their work.

Critics also find that the indirect uses of tests - their use by local educational authorities, as part of accountability procedures, to guide teachers in choice of topics and skills taught - impose undesirable limits on what teachers can do, and also on what schools and school systems can do. At the conference Ross Taylor, mathematics supervisor for the Minneapolis Public Schools, described how the Minneapolis system sought to improve its showing at the intermediate level on a mathematics achievement test, where students were several months below national norms. Analysis of the test showed that of 40 computation items, 15 were devoted to fractions. A crash program concentrating on fractions was instituted and scores went up. However, as Taylor notes, with the advent of the handheld calculator and the forthcoming change to the metric system, there are better ways to improve the mathematics curriculum at the intermediate level than singling out fractions for more intensive study.

Indirect use of tests can limit teaching in another way. The tests may be prescriptive not only of goals, but also of methods. A teacher, school, or school system seeking to build a curriculum based on discussion, primary sources for social studies, and firsthand observations for science might find itself handicapped when it came time for testing. There might be gains from such teaching in terms of students' feeling responsible for their own education or coming to understand how inquiry is conducted, but such gains are not likely to show up the next time a published test is administered. For immediate results on published tests, the premium approach is through use of recitation and textbooks.

4. Tests are too narrow in scope to provide for fair evaluation of new approaches to teaching.

Evaluation is an important part of the present effort to improve education, because without full evaluation an educational experiment loses most of its meaning. But critics maintain that the narrowness and inflexibility of published tests with regard to curriculum make them unsuitable for evaluation of new and potentially valuable approaches to teaching. In a recent review of government-sponsored evaluation, House and his colleagues (1978) question whether any battery of published tests can be used to evaluate large-scale social programs and they add that such tests have particular drawbacks when used with young children. Looking to the future, Porter and his associates (1978) call for abandoning use in evaluation of "standardized indices" because they can be employed without any knowledge of what one is supposed to be measuring; they ask instead that the connection between an innovator's goals and the evaluator's measures be made explicit.

II. RECOMMENDATIONS

We have examined briefly the major elements of the current situation in testing. Educational tests are now predominantly used for four purposes: accountability, selection, evaluation, and classroom guidance. Problems in each area can be identified in the context of contemporary criticisms of tests: Selection procedures are not completely fair to minority students. The use of tests for accountability is imperfect, and now new tests for accountability are urged. The tests are not a positive force in classroom teaching and, in some regards, are perceived as inhibiting an constrictive. Finally, the tests are not broad enough in scope to allow for fully satisfactory evaluation of educational programs. These kinds of criticism of tests and testing came up repeatedly at the conference and reflect much contemporary debate outside the conference (see, for example, the report in the Appendix).

We now turn to the main body of this report, discussion and recommendations on the ways that the conferees believe testing, and thereby education, can be substantially improved through research and development and other related activities. *These recommendations fall into three areas: better fitting of testing to the cultural background of students, better fitting of testing to educational objectives, and combining testing and teaching.*

A. Better Fitting of Testing to the Cultural Background of Students

Though there has been talk about the ideal of a culture-fair test since the beginning of testing, no test has yet been constructed which meets this ideal. Early expectations that perfect "fairness" or perfect standardization could be achieved by statistical means have proved unfounded. Under pressure from minority groups, test makers more recently have made some attempts to modify the language used in test instructions and test items so that it is more representative of vernacular language varieties. While these efforts are useful, it is by no means clear that a single test instrument can be equally representative of the language patterns of all major cultural groups in the population. Nor is it sufficient to modify surface linguistic features of tests.

Recent advances in our knowledge of language interpretation and of cognitive processes reveal that culturally different experiences and background knowledge may affect test performance in complex and subtle ways. Take tests of reading skills for example. Investigators at the Center for the Study of Reading (Steffensen, Jogdeo, and Anderson 1978) in well-controlled studies, showed the importance of the match between cultural background and reading passage content on reading speed, on reading comprehension, and on retention of information. Native American Indian children and majority culture children were given two stories to read about weddings. These stories had been carefully constructed to be comparable in vocabulary level and equivalent on measures commonly used as indices of reading difficulty, but one story dealt with wedding customs of Native Americans and the other with the "typical-American" wedding ceremony as depicted in magazines. Each group showed greater reading speed, superior recall, and better performance on comprehension questions for the story in its own tradition. This outcome seems commonsensical, but its implications for testing practices are profound.

While many items in achievement tests are designed to assess how familiar a student has become with material presented in the curriculum, other items use factual material or event descriptions in order to assess skills of comprehension, reasoning, memory, or problem-solving. Unless the material used for these purposes is equally familar to all cultural groups, differences in performance are uninterpretable. The difficulties of achieving "equal familiarity" in this sense are so formidable as to make the ideal of culture-fair tests appear unrealizable and perhaps, unreasonable.

An alternative strategy, and one adopted by many investigators specializing in comparative research, is to construct different forms of a "single test" - such that the difficulty level and knowledge domains assessed remain constant across forms, but the language and illustrative material are tailored to the specific background knowledge of particular groups. The aim here would be to devise "equivalent" culture-specific tests. This was, in fact, the strategy advocated by Binet, the inventor of the mental test. He rejected the notion that a uniform test could be used as a means of comparison of people from widely differing backgrounds and insisted that tests be appropriate to the background and everyday occupations of the individuals tested. Clearly the notion of perfect equivalence or comparability is also an "ideal" that can be more easily approximated for some types of tests than others. *But participants in the conference were of the opinion that recent advance in knowledge and technique make this a fruitful strategy to pursue.*

B. Better Fitting of Testing to Educational Objectives

Three of the major criticisms of current testing discussed above - inadequacy for accountability, negative or limiting effect on classroom teaching, and unsuitability for evaluating new approaches to teaching - derive either entirely or in considerable part from lack of fit between testing and educational objectives.

To understand how this lack of fit has come about, a look at the history of educational testing in this country is useful. The successful use of psychological and educational tests in World War I led to their wide adoption by school systems. Tests were used by the military services to select recruits for officer training and for training in various technical tasks. The method of sorting thus developed was the adapted to address the problem of sorting students in the civilian educational system. At that time, most students were not expected to finish high school and go on to college, and thus a major function of schools and colleges was to sort children and youths, encouraging only those who were most promising to go on. Much of educational testing and testing theory developed in this context of sorting. In this theory, validity is measured by correlation of the test results with some other relative measure like grades in school and college. An item is judged by its ability to spread scores out for sorting purposes rather than for its relevance to what the school is seeking to help the student learn.

In addition to these assumptions made to facilitate arranging students on a linear scale for sorting purposes, two other assumptions have tended to confuse and impede the improvement of educational testing. The first is the notion that the educational objectives of schools and colleges do not go beyond such simple skills as reading and computing and the recall of information in content areas. The second is the assumption that the attainment of important educational objectives can be adequately appraised by the use of paper-and-pencil tests alone.

We have now moved into an era where we seek to help all students achieve their full educational potential. As discussed above, we are attempting to use tests for a range of purposes much broader than sorting of students. Work on the curriculum by teachers, scholars, and textbook authors in the last two decades has made explicit a wider range of educational objectives. The coming of low-cost information-handling technology makes it possible for us to escape the limitations on testing imposed by 50-year-old scoring technology based on multiple-choice paper-and-pencil tests.

Our recommendations for achieving a better fit of testing to educational objectives fall into three categories. The current movement toward criterion-referenced testing should be strengthened. The potential of the new technology should be exploited in such areas as creation of computer-based test item pools, testing of problem-solving processes, and testing tailored to the individual. Better information must be developed and made available on tests and testing, including basic conceptions of testing, appropriate use of testing in education, and capabilities of existing tests.

1. Criterion-referenced Testing

A major response to the need for tests that serve purposes other than sorting has been the development of criterion-referenced testing. A criterion-referenced test determines what a student can or cannot do in a specified domain of educational objectives. Ideally, items are selected to give a proper representation of the domain. In contrast, traditional achievement tests, designed in the sorting tradition, eliminate test items that most students can answer, since these items do not produce the spread in scores desired for sorting. This latter practice tends to eliminate test items that represent what schools are trying hardest to teach and, as time goes by, may penalize better teaching by removing well-learned items in revised versions of the test. (These and other contrasts between criterion-referenced and ordinary achievement tests are discussed by Popham, 1978). Criterion-referenced tests can provide educators, parents, and others with a rather detailed picture of how well students, individually and in classes and schools, are learning in domains covered by the tests. Test makers may also develop data that allow comparison of performance in these domains among various categories of students and schools in different parts of the country. For example, the National Assessment of Educational Progress utilizes criterion-referenced tests and publishes national and regional performance data that schools can compare with their own.

Test makers are increasingly producing criterion-referenced tests. *However, the theory and practice of constructing and interpreting the results of criterion-referenced tests need further development.* Preparation of good criterion-referenced tests requires more careful analysis of the content domains being tested and preparation of more tests and test items than is the case with tests designed for sorting. Some tests presented as criterion-referenced are little more than reworked versions of sorting tests, without the requisite coverage of content. Construction of criterion-referenced tests may still be strongly influenced by the traditional objective of spreading out the distribution of scores on a bell-shaped curve.

Ideally, a school system should be able to give tests well-fitted to their chosen educational objectives at various grade levels and thus determine in detail how well these objectives are being attained and how this attainment compares with that of other systems. In addition to the barriers to achieving this ideal described in the previous paragraph, most test development has been confined to items testing simple skills or factual knowledge. The tendency to think of education in terms of the most basic skills has been intensified in recent years by the fact that some fraction of students still come out of school unable to read or to do simple arithmetic. This is a real problem for American education, but it is only one of a number of problems in a society that each year demands more highly skilled graduates from its educational system. The emphasis on our aspirations for universal literacy has tended to cause us to lose sight of broader educational objectives. Thus for example, many mathematicians, mathematics users, and mathematics educators today are concerned that the public and some educators are defining basic skills too narrowly, limiting attention to computation. Responding to this concern, the National Council of Supervisors of Mathematics (1978) published the following list of ten basic skills in mathematics:

Problem solving
Applying Mathematics to Everyday Situations
Alertness to the Reasonableness of Results
Estimation and Approximation
Appropriate Computational Skills
Geometry
Measurement
Reading, Interpreting, and Constructing Tables, Graphs, and Charts
Using Mathematics to Predict
Computer Literacy

Computation is an element in all of these skills, but the formal computational skills given greatest emphasis on most mathematics achievement tests, constitute only one of the ten basic skills listed. As mathematics teachers increasingly emphasize these other basic skills, tests used to assess their success in teaching must contain an appropriate selection of items in these other areas. Similar needs exist elsewhere in the curriculum. The more broadly conceived skills being called for in mathematics and elsewhere are often ones that reflect the needs of adults in analyzing and solving practical problems that confront them in their jobs and personal lives. Similar objectives - such as the ability to solve practical problems involving computation and reading - are found in a number of the competency tests being devised by the States.

Finally we note that although there are certain things that a school will wish every student to learn, we are also interested in encouraging students to develop special interests and capabilities of their own. Expecting every student to answer every question on a test is inconsistent with this goal and will tend to keep each student on a uniform path. A useful alternative (Zacharias*) is to make items on a test 1/3 mandatory, 1/3 choosable from a longer list, and 1/3 designed by the student to show his or her grasp of an idea or skill. As children move towards the higher grades of school their interests and skills diverge - as they properly should in a country that lists 20,000 different titles in its dictionary of occupational skills. Something other than uniformitarian testing is needed to supervise and encourage the diverse growth of children's competences.

2. Information-handling Technology

The fit between testing and educational objectives can be improved by taking advantage of the increasing availability of low-cost information-handling technology. We discuss three examples: computer-based item pools, computer-based testing for broader objectives, and tailored testing.

a. Computer-based Item Pools

The task of providing teachers, schools, and school districts with tests closely matched to their specific educational objectives can be made manageable at reasonable cost through the new technology. Central computerized pools of test items of varying complexity could be created by educational centers and test publishers. Users could be given access to these pools either through direct communication between the central computer and their local computer or indirectly through local information storage devices such as magnetic or video discs. Items would be indexed in such a way that users could assemble them to form tests suited to their needs. The system could also provide additional information on the items, including national error rates and comments by other users or cirtics. It would be easy to design direct access systems in which additional items as well as comments and additional error data on existing items could be entered by any user, but an arrangement in which the center acted as an intermediary would probably be better. Similarly, many teachers might desire the help of a local expert in compiling tests. Tests could be printed out for paper-and-pencil administration, but an eventual further refinement could be administration and scoring of the test via a personal computer for each student.

---

*References given without a date refer to papers in Part Two of this volume.

The statistical properties of a test as a whole are not, of course, the simple sum of the statistical properties of its individual items, and items do have interactions that are not fully foreseeable until they are tried out together in a test, but as new tests are developed, information on tests as well as on individual items could be developed. A teacher-assembled test could not have properties identical to those of current published tests; nevertheless, with so many local variations in populations, teaching styles, and curricula, school systems might conceivably produce new tests with greater predictive validity than current tests.

In addition to increasing teachers' choices in testing, computerized item pools would help to resolve the conflict between demands for complete access to all items and the need for secrecy of items. If the pool is large, secrecy of items in the pool is unnecessary. Only the selection of items for a particular test need be secret and then only before that test is given. Items used once can be used again. Indeed, part of the idea of the pool is to gather statistics as well as individual comments on items.

Mathematics teachers in Minneapolis have developed a city-wide testing program for selected aspects of junior-high mathematics that is suggestive of some features of a computerized item pool (Taylor). Teachers first decide on instructional objectives, then prepare tests keyed to those objectives, and finally write generation rules for each test item. A new form of the test is generated by computer each time the test is administered. Students use old forms for practice and as a way to learn the objectives of the course. Teachers change the objectives and the tests as they gain experience and improvements occur to them.

b. Testing Problem Solving

Technology now provides a means to improve instruction and testing in one of the ultimate aims of education - the ability to address more complex problems of the sort encountered in work and personal life. Such problems typically require us to bring to bear a variety of things we have learned in school or elsewhere; many are actually a series of problems, where each step depends on a previous step, and where various sequences of steps can be followed, some more efficient than others. Not only a student's answer, but also the efficiency of the solution strategy are of interest. Easily-graded standardized tests for these more complex problem solving procedures have been difficult or impossible to devise, so that a very important class of educational objectives has been left untested and thus undervalued. Information-handling technology now makes such testing possible at low cost. Computer-based testing can present problems realistically, can allow a student to proceed even when an arithmetical error is made, and can follow and evaluate the problem solving process.

THE NEW INFORMATION-HANDLING TECHNOLOGY

We have entered a revolutionary age in the technology of information handling. The microelectronic revolution, extensively documented in the September 1977 issue of Scientific American, is making possible a roughly tenfold decrease each five years in the cost of the integrated circuits that are at the heart of contemporary digital computers. There are available for individual use, at relatively low prices, sophisticated calculators and computers and means of generating images on video tubes:

Handheld calculators providing the four arithmetic functions and square root now sell for about $10, and the cost of calculators providing trigonometic and logarithmic functions has declined to the same level. Calculators able to handle 400-step programs are now at the $100 level.

Handheld instructional devices providing drill and games in arithmetic ($15-25) and in spelling, with a simulated human voice ($65), are selling well.

Personal computers which include a TV tube display for letters, numbers, and graphics; keyboard, processor, a sizeable memory, and weigh less than 50 pounds, are now available for as little as $600. Accessories include word-processing devices allowing flexible construction and editing of written text and production of typed copy. These computers are comparable in capacity to computers costing hundreds of thousands of dollars a decade ago, and are only the first of many that will become available at decreasing prices or with much greater capability at the same price. Incorporation of microcomputers into TV sets could make substantial computer power available in the home to each child.

A videodisc system coming onto the market for $700 provides the following, in conjunction with a regular TV set:

- 54,000 separate frames of full color picture image, or alphanumeric or computer information, on one side of a disc similar in size and cost to an LP record. The 54,000 frames correspond to 30 minutes of video at 30 frames per second, any part of which can be played at regular or reduced speed.
- Random access (dialed in more expensive models) to each of the frames, which are individually numbered and can be viewed individually for any length of time desired and may contain one quarter of a page of easily legible text (or a full page with special high resolution video tubes).

Intelligent videodisc systems under development include a computer which allows controlled sequencing of frames, which may be based on student responses, as well as the possibility of transfering computer programs from the disc to the computer. Video discs can store 10 billion bits of information (Encyclopedia Britannica contains 2 billion bits; a human chromosome has a capacity of about 20 billion bits, and the human brain perhaps 10,000 billion). This immense storage capacity could probably include on a single disc all the computer courseware ever published.

This use of technology is already well developed in medical education. For example, a test of skill in diagnosis begins with a statement of the patient's complaint. The student asks questions of the patient and orders laboratory tests. The computer then supplies the patient's answers and the laboratory results. The student can now render a diagnosis or continue to call for more information until ready to give a medical opinion. This testing procedure not only determines whether the student correctly identifies the patient's complaint, but also shows how the student goes about the diagnostic task - how he or she processes the information furnished by patient and laboratory, as revealed by the student's questions and the order in which they are asked. The results of several such sequences can help a student improve his or her diagnostic technique. This type of testing can obviously be applied in other areas. Investigation of process can be made as profound as one is willing to delve. "Some New Things to Test For" (page 19) lists some intellectual ploys used by scientists in wrestling with problems. Further examples of computerized testing are given later in this report.

c. Tailored testing

In a conventional test, the test taker works through a series of items in a fixed order, marking choice of response with a No. 2 pencil. One shortcoming of this procedure is that the test cannot measure accurately the abilities of test takers at the high and low ends of the score distributions for a heterogeneous group. In tailored testing, based on computer technology, the difficulty of the items is matched to the ability of the test taker. Multiple-choice questions appear on a display panel similar to a TV screen and the test taker indicates the answer on a typewriter-like keyboard. A correct answer is followed by a harder question, an incorrect answer by an easier one. No longer need the candidate waste time on items that are too hard or too easy, and responses to less appropriate items need no longer mar measurement based on more appropriate items. As the test proceeds, each response causes the computer to revise the estimate of the test-taker's ability. When the estimate reaches a specified level of reliability, the test ends (Urry 1977).

This approach is based on a "latent trait model" developed over 15 years ago (Lord 1952) that makes it possible to give different forms of a test, all of different levels of difficulty, to subgroups of a population of candidates and to obtain comparable scores, as though all had taken the same test. Recently, investigators have been developing tailored testing procedures for use by the Civil Service Commission (Urry 1977).

3. Providing Better Information on Tests and Testing

An essential element in fitting testing to the educational objectives of schools and school systems is the availability of reliable information on tests and testing. Educators, parents, policymakers, and the public need to know what tests can and cannot do, how testing can affect achievement of educational objectives, and what the capabilities of specific published tests are.

SOME NEW THINGS TO TEST FOR

Following is a partial list of intellectual tools and strategies that physical scientists and technologists often use in their professional work (Zacharias). We should be able to test students to see if they have, or are developing facility in the use of these and similar tools and strategies.

- Asking whether the problem is one that, in principle, can be solved.
- Asking what aspects of the problem can be ignored for the time being.
- Looking for a quick guess at the answer; looking for ways to improve the guess.
- Looking for ways to break the problem down into subproblems.
- Pushing an idea to the limits; pushing it to absurdities.
- Establishing upper and lower bounds.
- Bringing in outside information to bear on the problem.
- Recognizing and exploiting symmetries of the problem.
- Changing point of view and/or frame of reference to simplify the problem.
- Shifting attention between parts and wholes to develop new approaches to the problem.
- Anthropomorphism: What would I do it I were a ...?
- Seeking out clean approaches and solutions: taste, style, and elegance.
- Asking for help: Why, when, and of whom? Understanding that it is no disgrace to ask a question, or to be unable to answer one.
- Possessing a decent respect for multiple sources to evidence, as a basis for establishing firmness of belief.

a. Clarifying Basic Conceptions of Testing

The best known of all tests is the IQ test. The name IQ has been part of the common language for two generations or more. A person's score on an IQ test, expressed as a single number, has tended to be generally regarded as a precise, stable, objective measure of a centrally important characteristic of the person. This impression has persisted despite clear statements to the contrary by leading figures in test development, e.g., Edward L. Thorndike et al. (1927) who said the following:

> Existing instruments (for measuring intellect) represent enormous improvements over what was available twenty years ago, but three fundametnal defects remain. Just what they measure is not known; how far it is proper to add, subtract, multiply, divide, and compute ratios with the measures obtained is not known; just what the measures obtained signify concerning intellect is not known. We may refer to these defects in order as ambiguity in content, arbitrariness in units, and ambiguity in significance.

The common view of the significance of IQ scores has pervaded thinking about educational tests generally. The theory that has, explicitly or implicitly, dominated most thinking about educational tests is that they provide measures, analogous to length or weight, that can be used to arrange children, and what they have learned, in a simple linear order. This theory supported initial research on testing but it is obviously inadequate to our present conceptions of the complex nature of human knowledge. The present conference was not designed with the principal purpose of resolving issues in that debate. *We believe, however, that the time is ripe for a new look at these basic questions in testing. This new look must involve, in addition to leaders in testing theory, persons from various branches of cognitive science as well as teachers and scholars concerned with the substance of education. The results of this effort should lead both to new research on testing and to informational materials for teachers, teacher's colleges, parent-teacher associations, and the general public, setting forth as clearly as possible what tests can and cannot do, including issues for continuing study and discussion.*

b. Appropriate Use of Tests in Education

Tests are more and more widely used in the educational system for accountability, selection, program evaluation, and instructional guidance. The use of tests has often become a routine bureaucratic practice to which little thought is given. Tests designed for one purpose are often used for other purposes to which they are not well suited, and critics argue that tests often have perverse effects even when used as intended. *The conference believes a thorough program of studies of the use of tests in American education is needed to provide a basis for intelligent action. The studies should document the typical uses of tests by various educational agencies for various purposes and should investigate the effects of* *this testing on teaching, on the education of individual students, on educational innovations, and on budgetary and other decisions, and the appropriateness of these effects. These studies would make it possible to provide much firmer information on the suitability of various testing practices and how testing can be improved than is now possible.*

c. Information on Published Tests

There are a great many tests on the market today, and there is evidence that quality control is uneven. Further, because tests are kept secret from the public, there is little opportunity for individuals to challenge the findings of a low-quality test. *Our recommendation is twofold: formation of private consumer groups as sources of information about the capabilities of tests and provision for fully informative disclosure of test items.*

Consumer information groups. *Private independent groups should be formed, with private support, to provide test users easy access to information about the technical properties of published tests.* A single group will not be enough: there must be several that evaluate tests from their points of view rather than seek a consensus or minimal evaluation. The groups would explain from their perspectives how the tests were developed, including the domain of knowledge and skills from which the items were selected, the formulation of items and the procedures used to reduce bias. They would describe the rationale for the content and format of the tests and the procedures for scoring.

The model for such undertakings should not be the Underwriter's Laboratory, which approves the safety of an article, but something like Consumer Reports, which gives broadly accessible information about the factors of quality in an article and which then gives item-by-item information about how articles measure with regard to those factors. We are not recommending that tests be rank-ordered, or given ratings like "Acceptable", "Best Buy", etc. Currently, the Buros Mental Measurement Yearbook provide reviews of recently published tests, but these are given in a lengthy and detailed form that is probably not ideal for the unsophisticated consumer who wants to review what is available before choosing a test. We are not recommending that this activity be initiated by the federal government or a sponsored subsidiary such as an educational laboratory or center. Federal quality control over tests could evolve towards a national curriculum and in any event would be less desirable than several sources of test information developed through private initiative.

Fully informative disclosure of test items. *Students and parents affected by decisions based on tests should be able to see the tests - that is, the individual items and the student's answers - on which these decisions were based. This would seem at once a fundamental human right and a necessity, since testing procedures are fallible. The research community and any concerned layperson should have easy access to test items and the grouped responses of test takers so as to form independent judgments about the quality of the items and the presence of possible bias, and freedom to publish the items on which their judgments are based.* Arguments in a debate about an item or a test would seem to rest ultimately on appeal to items themselves, and without access to the items one cannot participate fully in the debate.

In considering the argument for secrecy, a distinction should be drawn between secrecy before administration of the test and secrecy afterward. Before administration, the topics to be covered may be revealed, but not the specific items and their answers. After administation, why should any matter be kept secret? Test publishers maintain a policy of secrecy after the test because the need constantly to redevelop old tests that had become public knowledge would increase the costs of test construction. One exception to the rule of secrecy is a federally-supported program, the National Assessment of Educational Progress, which releases 40 percent of its questions after each round of testing. Of course, most companies do release sample items which show some aspects of the questions asked. The computerized item pool described above requires no secrecy about items available for use on tests. We recommend also that the consumer information groups address themselves to the question of fully informative disclosure of test items.

C. Combining Testing with Teaching

Instructional guidance is the educational activity which is least served by existing tests. Yet the interaction between teacher and pupil is at the heart of schooling. Further, use of tests for purposes outside the classroom - accountability, selection, evaluation - should come out of classroom process, not be imposed on it like a foreign body. *Conferees envisioned, as a future ideal, testing merged into the teaching process and providing timely and rich feedback to the student, the teacher, and other interested parties as desired. The recommendations envision sustained development and research over the next decade or more exploring the use of testing in instructional guidance. Some approaches to this ideal already exist, but they are seen as only a beginning to be developed further and supplemented by other approaches.*

We see four elements as central to the development of this new combination of testing with teaching: full use of what cognitive science can tell us about the processes involved in learning and performing the intellectual tasks to be tested, exploitation of the new information-handling technology, strong participation of teachers and scholars in the subject areas, and good fit with the natural classroom situation. The contributions of each will be discussed separately, with the understanding that they must be brought together in the design of new teaching-testing systems.

1. Cognitive Science

Cognitive science today is a rapidly-growing interdisciplinary movement including psychologists, computer scientists, philosophers, linguists, anthropoligists and educators all converging on the analysis of human intellectual processes. An important component of this effort has been the detailed study of learning by students at all educational levels from preschool to college, coupled with a systematic effort to model partial and progressive states of knowledge using the computer. Any attempt to teach a subject involves some kind of theory or assumptions about what high-level performance in the area is like, as well as an explicit or implicit theory of the learning process. *Contemporary cognitive science is conducting a much more searching appraisal of both performance and learning than has heretofore been possible. The work in cognitive science is far more theoretically compelling than the "learning theories" of twenty years ago and, at the same time, much more closely tied to direct studies of learning in school environments. Development of better systems of instructional guidance will be greatly aided by this knowledge of what is involved in the student's progressive encounters with subject matter.* Presentations at the conference discussed efforts dealing with mathematics, reading, and writing, some of which will be described here.

One of the attributes of a good teacher is the ability to diagnose underlying misconceptions from a student's answers to a set of problems. The BUGGY computer program, developed by John Seely Brown and Richard R. Burton (1978), can undertake this task for certain aspects of mathematics instruction. The computer program includes "correct models", which represent the various ways to obtain correct answers, and "diagnostic models", which represent various ways that students typically obtain wrong answers. Like a good teacher, BUGGY is not limited in instructional guidance simply to indicating which answers are correct and which wrong, but can also indicate which of many misconceptions a student may harbor: the "bugs" in his procedures. There are many possible bugs in children's arithmetic (e.g., always subtracting the smaller digit from the larger: 1928-573=1455). Brown and Burton are devising a computer program allowing diagnosis of more than one bug, a task generally beyond the capability of a teacher.

It is tempting to assume that students make mistakes because they do not follow procedures very well: that the primary cause of error is simply inability to carry through a sequence of steps properly. But good teachers operate on the assumption, and the BUGGY program begins to demonstrate, as Brown and Burton note, that students are remarkably competent followers of procedures; the difficulty is that they often follow the wrong procedure. That is why cognitive science can aid good teaching by investigating not only the student's answer to a set of problems, but also the processes by which he or she obtained those answers, so that the correct processes can be learned.

Investigators are also developing psychological models of how children learn to read and write, based in part on comparative studies of how beginners and accomplished performers proceed. As described in a presentation at the conference (Collins), reading is a constant process of forming and correcting hypotheses, of which the reader is ordinarily unaware. Children sometimes fail to understand a passage not because they lack "reading ability", but because they lack some piece of knowledge needed to develop the correct hypothesis. Experienced readers face similar difficulties fixing the meaning of passages when they lack the necessary background knowledge or do not know the context in which a sentence is set. To give two extreme examples, if a reader is to understand the sentences,

> "the notes were sour because the seams were split" and
> "the house blew it",

the reader must know that one sentence concerns bagpipes and the other gambling activities.

Reading comprehension is necessarily intertwined with background knowledge: a passage must be about something. Students from different homes come to the reading task with differing background knowlege. It is possible to make progress in testing reading comprehension, rather than background knowledge, by relating the test to matters that have been specifically taught. Cognitive psychologists believe they can develop diagnostic procedures for aspects of reading analogous to those in the BUGGY program for arithmetic.

The Degrees of Reading Power test (Koslin et al) is based on recent work in psycholinguistics. The fluent reader is seen as using syntactic and semantic cues from the text, along with prior knowledge of language and content, to reduce uncertainties and confirm or disconfirm predictions concerning the meaning. A student can be considered to comprehend when he or she has eliminated all but the correct meaning of the text. The test items are passages of varying difficulty in which certain words have been deleted from the text. Subjects are asked to select the word deleted, from a list including several alternatives. Having carried out the comprehension process described above is a necessary and sufficient condition for choosing the right work. Persons taking the test can be allowed to inspect a list of topics in the item bank so that a reading test comprising items about which the subject knows something can be assembled, and possible effects of differing background knowledge can thus be reduced.

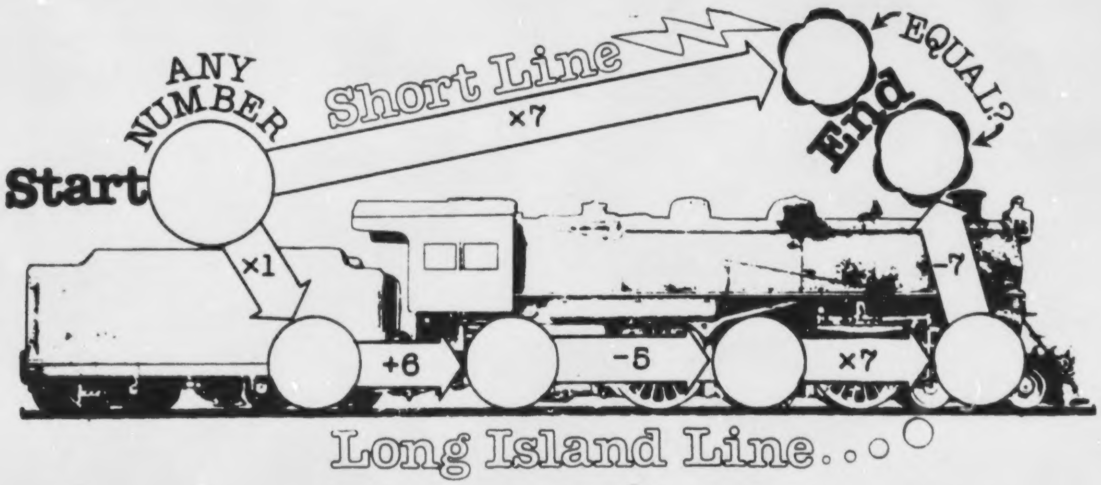
2. Interactive Teaching-Testing and Technology

Generally, human beings learn best when there is rich and immediate response to the learner's initiatives and when there is the possibility of branching, exploratory interplay between the learner and the teacher and testing is part of the interplay. Most schools are not able to provide that kind of learning environment because it would require something close to a one-teacher-one-student relationship. The rapidly increasing capacity and decreasing cost of electronic information-handling technology make interactive teaching-testing arrangements both feasible and economically attractive.

We have already described BUGGY, which represents an advance in cognitive science and also an exploratory effort in interactive teaching-testing, as is the example from medical education presented earlier. A number of other examples are provided by the PLATO system of computer assisted instruction which can provide graphic displays as well as words and numbers on a screen. A number of games devised for PLATO require solving mathematical problems. In How the West Was Won, a player is given three digits and must create a single integer using all three digits once and only once and using parentheses and the arithmetical operations +, -, and x not more than once each. Thus 4,1, and 6 could be used to produce (4+1)x6=30, or 6+1-4=3 among other integers. The integer chosen (and verified by the computer as correctly generated) moves the player's stagecoach or train as in a board game. Since moves have different consequences, the player should seek the most advantageous move by exploring the several integers that can be generated from the given digits and the consequences of the resulting moves. The student, playing against a fellow student or the computer, is thus exercised in both arithmetic and game strategy. Burton and Brown (1979) have devised a computer program that characterizes and evaluates a student's strategies in playing How the West Was Won and in other activities, proving a basis for improving them. Other PLATO arithmetic programs include two units on fractions, one involving the cutting of pizza pies and the other estimating the fraction corresponding to a point on the number line marked by a balloon which bursts when the estimate is a good one. These and other PLATO programs in arithmetic, in which a student is constantly tested in a non-threatening way, have proved to be engrossing and effective supplements to regular instruction.

Conferees proposed exploration of a new vision of a learning and testing environment extending what has been just described. In this vision of the future, school tests as we know them would cease to exist. The instrusive, specialized, institutionalized activity called testing would be absorbed into a new kind of learning and testing environment. Computers could accept inputs from students and teachers on an almost continual basis, extracted from the rich tapestry of ongoing learning activities. Instructional systems would accumulate an educational portfolio for each student, including a wide range of interrelated performance and situational descriptions. One would be as unlikely to cease all instructional activities in order to test a student as one is to stop conversing with a child in order to test his or her linguistic competence. Instead, testing would be a particular aggregation and analysis procedure applied to a continuously collected data base. Some of these aggregation would have an immediate impact on ongoing learning activities; other would be remote from the moment of data collection.

Interactive teaching-testing materials for schools can also be developed without a computer. As part of its system of instruction in computation and measurement for elementary schools, Project TORQUE has produced paper-and-pencil tests that provide immediate feedback to student and teacher. One type of test used is exemplified by the following figure:



Students choose a number and carry out the two branching sequences of operations. The test is self-checking: if the two sequences do not yield the same number, students know they must recalculate. These tests can be made public because the actual problem is not fixed until the starting number is selected. The same test can be used over again with a different number, selected by the teacher if desired. These self-checking tests have broader implications as well. Some students will be curious about why the tests work. What is it about the different paths followed that produces the same answer? Finding the reasons constitutes an introduction to algebra. These and other TORQUE tests can be categorized in terms of the specific skills which they assess. Consequently, they can be useful in demonstrating that students have mastered certain skills if required by school districts or state agencies.

3. Subject Matter

To construct good tests to asses what a student knows and can do in a domain of knowledge requires not only skilled test developers, but also outstanding teachers and scholars experienced in the content domain as part of the testmaking team. Persons with such background who look through collections of published tests in their area typically find in many of the tests numerous items that they regard as inappropriate exemplars of the knowledge area, as seriously ambiguous in wording, and too often as simply wrong. This leads them to conclude that knowledgeable persons were not involved in creation of the tests in question. Involvement of knowledgeable teachers and scholars in testmaking, important for any test, becomes absolutely essential in the development of testing-teaching systems and tests of problem-solving processes. The conference heard papers from teachers and scholars in the sciences, mathematics, and reading and writing which discussed educational objectives and intellectual processes in their areas with a richness and depth that only persons with their knowledge can bring to the testmaking process. Rough examples to be found in the present report include the outline of new things to test for on page 19 and the discussion of bugs in arithmetic processes on page 23.

4. The Natural Classroom Situation

The teaching-testing systems envisioned in this report are intended as aids to the teacher, rather than as substitutes for the teacher. Systems developed on this basis must take account of the natural classroom situation or be in peril of being rejected or poorly used. Developers should make full use of what is already known about the way instruction and instructional guidance take place in typical classrooms at the grade level and in the subject area in question and about the successes and failures of other newly-introduced systems. However, the new system envisioned is significantly different from what has gone before, and its reception by teachers and students is not predictable in detail. Research on these matters must be built into development projects and be used to improve the system.

The number of investigators interested in direct, intensive observation of classrooms is growing. They include anthropologists, psychologists, sociologists, and teachers with a talent for stepping back from their experiences and describing them. Conference participants working in this area provided several examples of how tests could be made more useful in the classroom. Test results closely related to the day's activities rather than to larger units are likely to be most useful to both the teacher and the student, and feedback on what probably went wrong would be even more useful. However, the need is not uniform across students. Teachers often believe they have enough information about many students in the classroom. For some students the teacher may feel the need for a great deal more information: The teacher doesn't understand the student, has "tried everything and doesn't know what to do." Testing procedures providing insight into such students' intellectual functioning could be very helpful to teachers. Both examples illustrate the more general principle that teachers, like other people, prefer arrangements that adapt to their needs rather then requiring them to adapt themselves to a rigid system. We do not wish to imply, however, that new systems should respond only to needs currently perceived by teachers. The most useful innovations may be those undreamed of prior to their invention, but they too must be tried in real situations and modified as necessary.

Conclusion

How are we to pursue this vision of testing merged into a teaching-testing system, fitted to the natural classroom situation, drawing upon cognitive scientists and teachers and scholars in the subject areas, and exploiting the rapidly developing information-handling technology? One way is to continue and perhaps expand support for research on classroom processes and human cognition and for development of new technologically-based testing and testing involving persons from subject areaa. Eventually, however, these points of view must be brought together if we are to have the working testing-learning systems envisionsed at the conference. Project TORQUE, with support from private foundations, is developing one such system in computation and estimation for schools, and the School District of Philadelphia (Charp), with support from local and Federal sources, has developed a different system based on behavioral objectives in various areas of education. We need to develop and experiment with more models. Projects to develop these models require research components if they are to achieve their full potential. Furthermore, development projects are often excellent sites for fundamental research in such areas as classroom processes and human cognition.

REFERENCES

Brown, John Seely and Burton, Richard R. (1978) Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science 2:155-92. (Reproduced in Part Two of this volume)

Buros, Oscar K. (1978) The Eighth Mental Measurements Yearbook, 2 vols., Edison, N.J.: Griffon Press.

Burton, Richard R. and Brown, John Seely (1979) An investigation of computer coaching for informal learning activities. International Journal of Man-Machine Studies 11:5-24

House, Ernest R.; Glass, Gene V.; McLean, Leslie; and Walker, Decker F. (1978) No simple answer: critique of the "Follow-Through" evaluation. Harvard Educational Review 48:128-160.

Lord, F.M. (1953) A theory of test scores. Psychometric Monograph No. 7.

National Council of Supervisors of Mathematics (1978). Position paper on basic mathematical skills. The Mathematics Teacher. 71:147-152.

Popham, W. James (1978) The case for criterion-referenced measurements. Educational Researcher 7:6-10.

Porter, Andrew C.; Schmidt, William H; Floden, Robert E.; and Freeman, Donald J. (1978) Practical significance in program evaluation. American Educational Research Journal 15:529-539.

Steffensen, M.S.; Jogdeo, C.; and Anderson, R.D. (1978) A cross-cultural perspective on reading comprehension. Technical Report No. 97. Champaign, Ill.: Center for the Study of Reading.

Thorndike, Edward L. et al. (1927) The Measurement of Intelligence. New York: Teachers College Bureau of Publications.

Urry, V.W. (1977) Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement 14:181-196.

Note: References in the text given without date refer to papers in Part Two of this volume.

INDIVIDUAL COMMENTS

Two participants asked that the comments below be appended to this Report.

Comments by Donald Ross Green, CTB/McGraw Hill:

The report begins by noting that current standardized achievement tests are widely used and that they must therefore be considered useful by many people. The report then proceeds to make criticisms of these tests both in the section so labeled and thereafter without noting either their merits or that many people do not accept these criticisms as reasonable or as based on facts. Perhaps the merits of these tests are indeed so evident they need not be reiterated but certainly the alternate points of view about at least a few of these criticisms should be cited. This commentary will therefore give a brief rebuttal to three of the criticisms and then try to reinforce the most cogent point in the report.

Curriculum Fit

A truly difficult issue is the matter of the fit between the curriculum and content of the test. There are substantial differences among the major achievement test batteries in what skills are emphasized and at what grade levels these skills are tested. Thus school systems typically have a test selection committee which includes classroom teachers and curriculum supervisors. These committes pick the tests they believe best, usually emphasizing the fit of the test to the curriculum of the system. (The Center for the Study of Evaluation at UCLA has prepared forms and procedures to assist people in studying tests this way.[1]) However, it is common to find people in a school system who disagree with the curriculum specifications of their system and therefore with the test selected. These people and others often do not follow the curriculum specifications closely in either program planning or instruction. Partly for this reason and partly because of the vested interest of the larger state and national communities in the outcomes of education, some evaluation experts do not endorse program evaluations which use tests specifically designed just to fit a particular local curriculum. They argue that a good evaluation is broad and tells one about student learning in the areas not emphasized as well as in those that are.

Bias

The most emotional issue of testing concerns racial and ethnic bias, especially in the context of selection. Although the conference was concerned with testing of achievement not aptitude, most of the statements about bias refer to selection and do not necessarily apply to achievement tests. The evidence about achievement test bias does suggest that there is some bias in standardized tests, that this amount is probably small in most skill areas for most minority students, and that there is probably more bias in unstandardized measures. Almost all the currently used standardized achievement batteries have been through editorial and empirical study leading to revisions designed to reduce this bias. Nonstandardized tests lack the data, and other methods of assessment lack the objectivity that permit one to study this matter, facts which preclude any serious effort to either identify or reduce such bias.

---

[1] Hoepfner, R. Achievement test selection for program evaluation. In Wargo, M. & Green, D.R. (Eds.), Achievement testing in minority and disadvantaged students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill, 1978.

Since education is a process of enculturation, a prerequisite for the complete elimination of bias is a truly multicultural curriculum and program. However, this is a distant goal and as long as there are cultural differences among groups in this society that are not fully understood or accepted by all concerned, some cultural bias in the measurement of achievement will occur. In some instances the bias will be against the majority group but logically it will more often be against minorities.

Test Security

It is averred that standardized tests are kept secret. The basis for this assertion seems to be the fact that school systems do not like to let students study the specific answers to the questions on a test ahead of time. This position is reasonable and does not preclude student practice on similar test items ahead of time. In fact all published tests provide some practice as part of the test. However, it may be desirable to provide unskillful students with substantial additional opportunities to practice taking test items of the sort to be given. Many teachers do this and some school systems do this regularly.

School systems who buy achievement tests are free to distribute copies of the test before or after its administration to teachers or others as they see fit. In fact many standardized test reports provide information about student responses item by item. These data can be reviewed with students and/or parents in conjunction with the item descriptions usually provided or with the test booklet itself. Most teachers appear to prefer the first of these two procedures and so do most school administrators.

Frequently evaluation designs call for both fall and spring testing with the same test because this procedure provides the most accurate measure of growth. Item by item reviews with students in the fall are clearly inappropriate here as is also true when the school intends to use the same test booklets the following year.

Interactive Teaching-Testing and Technology

Finally this report properly asserts that there is limited value to the traditional norm referenced achievement test for the guidance of classroom instruction. Not only are they given only once (occasionally twice) a year but they are not designed to perform this task as their principal function. Therefore aside from providing useful feedback about progress on a year to year basis they are of relatively little help to teachers and students. To the degree these tests provide detailed criterion referenced information (which some of them do) they may also be useful for planning classroom instruction; but by and large they are much more useful to curriculum planners, school administrators, and others who have to make school or system-wide decisions.

Better procedures for helping teachers plan and execute instruction have long been needed, although the many criterion referenced tests now in use are a step in the right direction. I support the recommendations of this committee concerning the use of our growing knowledge of cognitive processes and instruction in conjunction with computer technology to provide ongoing instructional guidance to students. One caveat is needed: school learning is primarily a social process and any machine-oriented procedures which do not fully recognize this will prove ineffectual.

I also would note pessimism about any short term (e.g., less than ten years) possibility of using the data so collected to replace that now provided by standardized tests for assessing group status and large-scale program evaluation. Reasonable procedures for making use of these data do not exist.

Comments by Jerrold R. Zacharias:

I. The report called "Testing, Teaching and Learning" suggests a consensus when indeed there was none.

II. a. Some of the participants believe that meaningful numerical scores cannot result from any tests, good or bad; cannot be applied to people of any sort, for any purpose or by any means whatsoever.

b. Some participants believe the opposite of this and some waffle as did the report.

III. Some of the participants believe that cognitive science, except for the wisdom of the ages, is not yet in shape to be applied to schools. The report says the reverse.

IV. Everybody belongs to some sort of minority group, with special abilities, disabilities and interests. The report refers to the present standard "Minority Groups".

V. The emphasis on the various roles for computers in the suggested "Computer-based Item Pools" results in overlooking the hard work by sophisticated people needed to make and to test test items and tests.

VI. Inadequate reference is made to the present scandalous inadequacy of the subject matter which passes for mathematics and language learning.

VII. No decent research on tests can take place if it depends on commercial tests as they now exist. This leaves NIE with the need to violate its own mandates before research gets started. No tickee, no research.

PART TWO: INVITED PAPERS

I. INTRODUCTION

Educational Objectives and Educational Testing 36
RALPH W. TYLER

Technology 48
ARTHUR S. MELMED

II. EDUCATIONAL OBJECTIVES AND COGNITIVE MODELS

The Sciences and Technologies 54
JERROLD R. ZACHARIAS

Mathematics
ROBERT B. DAVIS 60
LEON HENKIN 83

Mathematics: A View from the Schools 98
ROSS TAYLOR

Information Processing Analyses of Mathematical Problem Solving 113
JOAN I. HELLER and JAMES G. GREENO

Reading 136
ALLAN COLLINS and SUSAN E. HAVILAND

Writing 146
CARL BEREITER

A School Perspective on Language 167
PARKER DAMON

Automated Dictionaries 180
GEORGE A. MILLER

III. TESTING

Some Emerging Trends in Testing 186
NORMAN FREDERIKSEN

Cultural Considerations: African-American 204
ASA G. HILLIARD, III

Cultural Considerations: Hispanic-American 219
AMADO M. PADILLA

Instruction and Testing in the Philadelphia Public Schools 244
SYLVIA CHARP

Project TORQUE 261
JUDAH L. SCHWARTZ and EDWIN F. TAYLOR

Diagnostic Models in Basic Mathematical Skills 273
JOHN SEELY BROWN and RICHARD R. BURTON

Effectiveness Measure in Reading 311
BERTRAM L. KOSLIN, SANDRA KOSLIN and SUSAN ZENO

I. INTRODUCTION

EDUCATIONAL OBJECTIVES AND EDUCATIONAL TESTING: PROBLEMS NOW FACED

Ralph W. Tyler
Director Emeritus
Center for Advanced Study in the Behavioral Sciences
Senior Consultant, Science Research Associates

Criticism of educational testing, which began as an academic debate in the Sixties, has become a public controversy today. The National Education Association has called for a moratorium on testing while at the same time many states have adopted new minimum competency testing programs. Confusion has reached a point of near crisis. It is high time to review the progress and problems of educational testing and seek to develop constructive policies to improve educational evaluation.

BACKGROUND

Assessment of educational achievement has been practiced for several thousand years. The prevailing name for this activity changes from time to time -- examining, quizzing, testing, measuring, evaluating, appraising, and, currently, assessing but the primary function of ascertaining the educational attainments of students has remained constant. The scientific or systematic development of assessment has largely taken place in the twentieth century. During this relatively short period of time, some profound changes have taken place in both the purposes and expectations for testing.

The successful use of psychological and educational tests in World War I led to their wide adoption by schools and other civilian institutions. When America drafted two million men for military service in 1918, the problem of organizing and training this large number of persons who had had no previous military experience was overwhelming. Who were to be selected for officer training and who for the variety of technical tasks - construction battalions, signal corps, quartermaster corps, and the like? The psychological advisors developed the Army Alpha Test and other classification tests that provided the basis for selecting and classifying this large assortment of young men. After the war, group tests, both of intelligence and achievement, were constructed and developed for school use, employing the same methodology that was formulated for the Army Alpha and other military classification tests.

This methodology is designed to arrange those who take the test on a continuum from those who make the highest score to those who make the lowest. Such an arrangement permits one to identify the position of any individual in terms of his standing in the total group. For example, Mary Moore is at the 98th percentile, and Richard Roe is in the bottom quarter of the group. Furthermore, the arrangement enables one to identify those who compose any desired fraction of the group; for example, the persons in the top quarter or those in the lowest third.

By administering the tests to a representative sample of a defined population, like children in the third grade of U.S. schools, the continuum on which those test scores are arranged is the distribution of the scores of all American third graders. This makes it possible to overcome the limitations in comparisons within an individual classroom, school, city or state by referring to a national norm. The original purpose was to sort students or grade them from excellent to poor. Because the problems that were then recognized were those of developing objective, reliable measures of complex human behavior patterns, primary attention was devoted to methods for constructing so-called objective tests - true-false, multiple response, short answer forms - and techniques for maximizing reliability, and constructing scales of equal units of measure. Questions regarding the appropriate purposes for assessment, what educational objectives need to be assessed, and what techniques of appraisal and analysis of data are consistent with the purposes, were not given significant attention.

This briefly describes the conception of educational testing that prevailed during the twenty years following World War I. It is still the dominant view of psychometricians in spite of the work done in the evaluation of the Eight-Year Study (1933-41) in which new conceptions and new procedures were developed. Now, however, the changing emphasis in schools and colleges from sorting students to educating them, combined with the variety of uses to which educational tests are now put, are raising serious questions regarding both the adequacy of earlier test theory and the validity of the assumptions on which it was based.

EDUCATION CONTRASTED WITH SORTING

Many of the educational doctrines and practices in the United States were developed largely before 1945, and were applicable to the structure of society and the characteristics of the clientele of earlier times. In our nation's early years, not all children were expected to succeed in school tasks, nor did they actually need to succeed so far as making a living was concerned. This was a time when most people were employed as unskilled or semi-skilled laborers and only 5 percent were in professional or managerial occupations. Only a few persons, therefore, were able to utilize a college education in their work, and most people could survive with little or no education.

Under such conditions, a major function of schools and colleges was to sort children and youth, pushing out those who were judged least promising for further education and encouraging a few to go on. The lock-step progress of instruction and the grading system were developed to sort students rather than to help every child and youth get an education. By moving the whole class at the same rate from topic to topic, pacing the movement in terms of the performance of the average students, those with more difficulty in school learning would be certain to get farther behind and most of them would give up trying to learn. The resulting feeling of inadequacy was reinforced by the grading system often supported by tests designed for the same purpose, which gave low marks to those having difficuty. At the same time, those who mastered school tasks easily and quickly were assigned high ranks which encouraged them to go on with their formal education.

Those grading, testing, and sorting policies and practices have existed for so long that we rarely note how sharply they differ from those of an institution devoted wholly to teaching and learning. For example, if you or I want to learn to play golf we go to some institution such as a golf club that employs a "pro" whose job is to teach us. We don't expect that after a few practice periods he will say, "You are getting a D in your work. I may have to fail you, if you don't improve." Instead we expect him to say something like this, "You are making progress on your drive, but you need to bring you full body into the swing. A little later, I'll give you some more practice on your putting, to increase accuracy and decrease power."

Our schools and colleges have been only partly concerned with helping each student learn, and they have devoted a great deal of attention to grading, classifying and other sorting functions. Most educational tests were developed to serve these purposes. Then education was rationed and sorting was a means of deciding who got educated. Today with the use of science and technology unskilled labor represents only about 5 percent of the U.S. labor force. A young person without education finds very few jobs available to him. The critical task of the school is no longer one of sorting students but rather one of educating all, or almost all, young people to meet the needs of modern society and to help them take advantage of the greater opportunities now available.

VARIETY OF PURPOSES FOR WHICH TESTS ARE USED

While the function of sorting students is waning and the need for so-called norm-referenced tests is diminishing, the effort of schools to educate all children is creating a demand for tests for a wide range of educational purposes. In the classroom some teachers are using tests to make needs assessments in order to individualize instruction. Some are using tests as means for placing students in the sequence of an educational program. Some monitor the progress of individual students by periodic testing. Individual students are taking mastery tests to demonstrate their successful completion of a unit of instruction. Tests are being used by many teachers to diagnose the learning problems of individual students who are having difficulty.

Curriculum builders are using tests for the assessment of group needs. Administrators are using tests to monitor the progress of classes and schools. Administrators are using tests to assess student achievement at certain age or grade levels. Tests are being used to evaluate educational programs. Tests are also being used by various persons to appraise the educational effectiveness of instructional equipment and materials. Tests are being used by guidance personnel to furnish information to the student that is helpful in his planning for the future. Tests are also being used by legislators, and other officials at various levels of government to obtain information about the progress and problems of education in their jurisdiction. Tests are also furnishing information to the general public about the progress of education.

These varied uses of tests were never anticipated when early test theory was being developed. Most of psychometric theory is based on the notion that educational tests are primarily to be used for sorting. In terms of this theory, validity is measured by the correlation of the test results with some other relative measure like grades in school and college. An item is judged by its discrimination not by its relevance to what the student is being helped to learn and under what circumstances he can be expected to use what he has learned. A test is also judged in terms of its internal consistency, thus assuming that things learned are highly intercorrelated.

In addition to these assumptions made to facilitate linear measures of psychological functions there are two that continue to confuse and impede the improvement of educational testing. One of these is the notion that the educational objectives of schools and colleges are chiefly skills, like those of reading and computing, and recalling information in the content subjects. The other is the assumption that the student's attainment of the important educational objectives can be appraised by the use of paper-and-pencil tests. It is increasingly being recognized that neither of these assumptions is tenable.

EXAMPLES OF EDUCATIONAL OBJECTIVES EMPHASIZED TODAY

An examination of curriculum materials designed for use in schools and colleges and a review of professional literature dealing with the teaching of the several school subjects quickly indicate the fact that in every field of study, several major objectives are commonly recognized and the instructional programs are planned to help students attain these goals. In science, for example, the following kinds of objectives are among those recommended:

1. To understand certain basic scientific concepts and generalizations and use them in observing and explaining natural phenomena.
2. To carry on inquiries seeking to understand puzzling natural phenomena.
3. To make reasonable interpretations of data about natural phenomena obtained from experiments.
4. To know and use dependable sources of information relating to science.
5. Skill in the use of scientific instruments and other apparatus.

In mathematics, the following examples are illustrative:

1. To recognize questions and problems requiring quantitative relationships for understanding them or solving them.
2. To compute appropriately and accurately.

3. To use common algorithms appropriately in dealing with the affairs of daily life.
4. To understand basic concepts related to the number system and measurement.
5. To use approximate estimations when they are sufficent and efficient.

In social studies the following are examples:

1. To understand the functioning and operations of a modern political system; a modern economic system and one or more social systems.
2. To use historical methods in seeking to understand the development of a contemporary social institution and a major social problem.
3. To understand the continuing tension in a democratic society created by the belief in balancing values such as freedom and responsibility, equality and excellence, self-interest and altruism.
4. To recognize serious social problems and analyze them objectively.
5. To predict the possible and probable consequences of courses of action proposed to solve major social problems.
6. To use dependable sources of information in the social studies.

The following are illustrations from the fields of writing and composition:

1. To select and organize content to be included in a written communication to a specified individual or group for a specified purpose.
2. To write an exposition that explains the subjects in terms that are clear to the intended recipients and are likely to hold their attention.
3. To write a description that describes the subject in terms that are clear to the intended recipients and are likely to hold their attention.
4. To follow accepted conventions of writing including spelling, punctuation, capitalization, sentence and paragraph structure.

In reading and in courses in literature such objectives as the following are found:

1. Decodes printed material correctly.
2. Uses appropriate methods of attacking and decoding words new to him.
3. Comprehends the plain sense of the material he reads.
4. Interprets material he reads by bringing to bear other relevant ideas and information.
5. Recognizes the intent of the author and judges the probable validity of the statements made.
6. Enters in sympathetically to the vicarious experience provided by literary works.
7. Makes appropriate emotional responses to poetry, drama, and novels.

Clearly, when educational programs seek to help students learn such a variety of understanding, skills, approaches to problems, and appreciation, the notion that a single test score can appraise either the program or the student, is an absurd conception.[1] It is also obvious that a paper-and-pencil test is unlikely to indicate the attainment of some of these goals.

HOW TO IMPROVE TESTING

Among professional groups there is wide recognition of the inadequacy of the commonly used tests for the range of purposes for which tests are now needed but there is less consensus about what can be done. One proposal frequently made, is that of reviewing the very large collection of test items in the repositories of the major test publishers to identify those that are appropriate for each of the different purposes. This proposal overlooks an important principle of all testing namely that a test is a representative sample of a defined universe. No item is in itself a dependable test. An example may serve to illustrate this point.

A test for "comprehends the plain sense of the material he reads" to be used with sixth grade students should be composed of a representative sample of the universe defined as "the material that he could be expected to read." This sample could be based on stratification in which case separate random samples would be obtained of such strata as, school books used in the middle school, stories available in the library, directions for assembling appliances and equipment used by sixth graders, newspapers they read, etc. These strata would be put together with proper weighting to form representative samples of the universe for testing. Each item will probably vary from others in vocabulary, in sentence structure, in appeal, and in other ways. No one item can be taken out of it as a proper test because it is not known how well it reflects the defined universe of "material he reads." Furthermore, part of the variability comes in the "appeal" of the item, that is, his sense that it is meaningful and hence evokes his effort to respond. We have been using tests with middle class children most of whom have been taught that answering adult questions is important even though they seem silly or in error. Many other children have not been taught to try hard on things that seem unimportant to them. Hence, there is a specification in the statement of the objectives, "material he reads," that makes it different from a sample of all reading material.

To improve testing we must develop tests appropriate for the purposes and not simply assemble old test items that were not designed for the objectives or for the purposes for which they are used. Along with the improvement of the tests themselves, those who use them need to be helped to understand what kinds of tests are required for each purpose.

---

[1] Bloom's Taxonomy of Educational Objectives indicates this in detail.

PROCEDURES FOR BUILDING AN EDUCATIONAL TEST

In the construction of better tests, earlier theories and assumptions that are no longer tenable must be discarded and test constructors can be guided by the basic principles of sampling and the increasing knowledge of complex human learning.

Identifying and Defining Objectives

The first step is to idenfity and define each major educational objective that the student is expected to attain. This is not difficult for educational programs that seek to help students develop certain skills like adding whole numbers or using a compound microscope. But in content subjects like science and social studies, test makers have often identified only the content covered and have not defined what the student is to do with the content. Hence, the test items usually require the students to remember statements of the content, not often asking him to explain, to predict, to solve problems, etc. Yet these are things today that are usually emphasized by teachers in these subjects.

The examples of objectives given earlier in this paper illustrate the variety of things that a student may be expected to do with content, but they are general objectives requiring additional specification to guide test construction. However, the specification should not be in conflict with the fact that education is always aimed to help students generalize, that is, to be guided by principles, modes of approaching situations, cognitive maps, and the like rather than by rigid rules and habits. The objective should be clearly defined at the level of generality intended by the course and/or the teacher. The current fad of stating objectives very specifically may be useful for training unskilled or semi-skilled workers but it is not appropriate for educational objectives emphasized by schools and colleges.

The definition of what the student is expected to learn will be broader for a test to be used at the completion of a learning sequence, such as a course, to ascertain whether the student has learned what the course seeks to teach than for a test to be used at the conclusion of a unit of instruction to find out whether the student has learned what the unit was designed to help him learn. For a test to be used to place students appropriately in a sequence of instruction, the definition of objectives will indicate the essential things, concepts, skills, and the like which the student should have learned in order to proceed successfully in the next unit of the sequence. In brief, the definition of objectives should state what the student is expected to have learned for the particular use to which the test is to be put.

Identifying and defining objectives require major participation by those in the subject field who have experience in teaching the subject and in constructing courses and other educational programs. The role of test constructors is that of questioning and asking for examples in order that the definitions are clear and can guide test construction. The new curriculum projects of the past twenty years have formulated a number of educational objectives different from the emphasis on memorization. They are helpful sources for identifying objectives. The new technologies being used in education such as self-paced instruction, computer assisted instruction and other forms of programmed procedures have required the designers to define the objectives of these programs. Current work of behavioral scientists on cognitive development, suggests some possible educational objectives. These are useful sources in defining objectives for testing but the important criterion of an objectives for an educational test is the intention of the educational program to help the student attain this objective.

Identifying Situations Where Objectives Are Utilized

A second step in constructing an educational test is to identify the situation in which the student can be expected to utilize what he has learned. The logic of this step should be clear. If one has learned something, he has internalized it and can be expected to use it whenever it is appropriate. Hence, if we wish to find out whether a student has learned something we should look at those situations where the learner can use what he has learned. A test, therefore should sample these situations and furnish an indication of whether the educational objective has been attained.

Note that this general formulation of a test does not limit the situations to those where the student is presented with written questions and makes a response on paper. Arithmetic skills, for example, may be shown in plans for a table to be constructed, in making out a budget, and in estimating the amount of fuel required to go to a fishing lake. Or, to take another example, understanding of certain principles of science may be shown by plans for a flower garden, together with an explananation of the choice of certain flower seeds, certain soil, method of soil preparation, and so on.

Devise Ways of Presenting Situations

A third step is to devise ways of presenting these situations to the student that will evoke the reactions that the normal situations would evoke or alternatively, devise ways of observing practicably the student's reactions in the situations that normally arise where the attainment of the objectives can be shown. It is often difficult to devise artificial situations in which the simulations seem so real as to assure the student's motivation to respond. This is the reason for using the term "evoke" rather than the more passive phrase, "requiring a response." As was mentioned earlier, some children, particularly those from middle class homes, take formal test situations seriously and try to respond. Lower class cildren, on the other hand, usually come from backgrounds where tests are rare and they appear to be silly games. Tests that use artificial situations need to be reviewed carefully and tried out to ascertain whether they evoke the kind of appropriate response from the several kinds of students tested.

Some of the new technological devices can produce some simulations that give much of the illusion of reality. Videotapes, and video discs, can present situations that can involve the student in the action so fully that it evokes the relevant things he has learned. Computer programs have been developed for testing the ability of the student physician to diagnose the illness of a patient for whom that data are in the computer's memory so that he can obtain specific information as he seeks it in his effort to make the diagnosis.

When it is possible to obtain representative samples of the student's reactions in real situations these should be used as tests. Trained observers can obtain a wealth of information about what students have learned. Teachers can collect this kind of test data, and can be helped by training to get representative samples of student reactions. These situations should not be overlooked in constructing tests.

Devise Ways of Obtaining a Record

A fourth step in constructing an educational test is to devise ways of obtaining a record of the significant aspects of the student's reaction in the test situation. In most written tests the student makes the record by writing his response or by checking one or more of the responses presented to him. The popularity of the multiple-response test is due to the ease of recording the responses as well as the apparent simplicity of appraising them. When the skill or ability is accurately indicated by a product the student makes, such as a composition, a dress, a work of art, the product itself becomes the record which can be preserved for careful appraisal. Observation check-lists, anecdotal records, video-tapes, audio-tapes, and photographs can furnish records of reactions that do not furnish their own record, as in speaking, social interaction, skills in using laboratory equipment, and skills in playing games.

New technologies have furnished new and much needed means for recording behavior. The slow motion picture records movements normally too rapid for observation. This has greatly improved diagnosis in physical education. The audio-tape recorder is a great asset in appraising oral performance, musical skills, and group problem-solving. Video taping provides a means of recording both behavior accessible to sight and oral behavior. Computers have tremendously extended the practicable ways of recording arithmetic skills, decision processes, and many forms of choices from multiple lists. No longer can the use of traditional paper-and-pencil tests be defended solely on the ground that there are no other practicable possibilities.

Decide on the Terms to Use in Appraisal

A fifth step is to determine the terms or units in which the student's reactions are to be described, appraised or measured. The age-old tradition of scoring or grading tests in terms of the number of correct responses has often prevailed when the sum of correct responses does not furnish a useful or a reasonable accurate appraisal of the student's attainment of the objectives. An important consideration is to use terms or units that properly reflect the desirable characteristics of the student's reactions in contrast to the undesirable or less desirable. For example, in appraising the reactions of a student in attacking a problem of resource allocation, the number of relevant major factors he considers can be the units of desirable characteristics, while the failure to work out their interrelation can be a descriptive term of an undesirable reaction. In some cases, as in diagnosis, the terms used may need to indicate syndromes or types of difficulties the learner is encountering. In appraising products, like compositions, two appraisal schemes are sometimes used - one indicating the level of quality of the total product and the other furnishing a report on the number of different desirable or undesirable features, like compound sentences and misspelled words.

Devise Means for Getting a Representative Sample

A sixth step is to devise means for getting a representative sample of the student's reactions which is adequate for the purposes the test will serve. We all know that human behavior may vary under different conditions, even when those conditions are clearly defined. One may read newspaper articles easily and find it hard to read directions for assembling an appliance. One can understand the scientific principles that help to explain photosynthesis and not understand the principles that help to explain the operation of an electric motor. One can use a large vocabulary in speaking of art and still be limited in the words he uses to discuss social issues. For a test to furnish reliable information about what the student has learned it must be based on a representative and adequate sample of the situations in which this learning can be exhibited.

To obtain a representative sample of something it is necessary to define the universe of this "something" and then to draw random or stratified samples from this universe. For example, to obtain a representative sample of the reading situations for a sixth grade test in comprehension, it is necessary to define the universe of things the sixth grader reads, such as kinds of text books, stories, newspapers, directions, etc. It is from this universe that samples can be drawn for testing.

Sorting tests are usually made without defining carefully the universe from which the items are drawn because their reliability is estimated from the correlation between two forms of the test. One of the reasons for the inability of the Panel on the Decline of the Scores on the Scholastic Aptitude Test to determine whether there had been a real decline in the vocabulary of the takers of the test (the greatest decline was on the vocabulary score) was the fact that the test constructors did not have any defined universe from which the vocabulary items were taken. It is quite possible that the samples were biased in the direction of words used some years ago that are now less frequently found.

A sample is likely to be representative if it is drawn from the universe by a random procedure or by being divided into strata each of which is randomly sampled and then put together with the samples from the other strata by the appropriate weights. The adequacy of the sample, that is, its required size, depends both upon the precision demanded for the purposes that test is to serve, and by the variability of the student's reaction to the different situations. The more variable the student reactions are the larger the sample required for the desired reliability.

Some of the new technologies help to solve the problem of obtaining an adequate sample of the student's reactions for a measure reliable enough for the purposes involved. A computer can keep in its memory the results of the student's performance on test exercises taken periodically until a relatively large number have been recorded and then reliable appraisals can be made from them. A teacher or other observer can code observable reactions and record instances whenever they recur by using the computer to enter into its memory the code for the student and for the kind of reaction observed. This makes possible comparing behavior over different time periods and over cumulative frequencies. It adds greatly to the reliability and precision of educational testing for many different purposes.

Following these six steps appears likely to produce tests that can helpfully be used for the range of purposes for which tests are now being used. These procedures are being followed now by a few test constructors and others can develop the necessary know-how as the demand for better tests increases.

LONG RANGE STRATEGY

What is proposed involves the development in detail of a comprehensive theory of educational testing and the production of many examples of practicable procedures in harmony with the theory. The strategy also calls for pilot programs of testing for each purpose thus exploring the problems and difficulties that may be encountered in developing better tests and more helpful testing procedures. These pilot studies should also furnish experiences that can validate or indicate needed modification in the theory and make it more useful in guiding practice.

This program of research, of test construction, and of pilot studies cannot best be done by psychometricians or test technicians. It requires participation and cooperation by highly knowledgeable scholars and scientists, curriculum builders, experienced teachers, psychologists and developers. It requires continuing cooperation with schools and colleges where tests, testing and educational improvements generally can be tried, observed and appraised. It can accomplish in the next twenty years what was done for sorting tests in the twenty years following World War I. We need this intellectual guidance in order to derive the potential benefits of testing as a tool of learning.

SHORT-RANGE TACTICS

As the long-range strategy is getting underway there are several constructive activities that can be pursued immediately. Encouragement can be given to the efforts of those competent, scholars and scientists who are seeking to identify the important contributions their subjects can make to the general education of young people since their work will lead to new or revised educational objectives. They can be helped to initiate or participate in test construction activities with teachers and others interested in improved testing. Similarly, those knowledgeable persons who are seriously seeking to use new technologies to individualize instruction and in other ways to improve education can be stimulated to work out the possibilities of these technologies for improving tests and testing. Those working at the cutting edge of the fields of cognitive psychology, social psychology, personality development and socialization can be encouraged to pursue the implications of their work for educational objectives and testing. As an aid to bringing these various activities to bear upon the improvement of educational testing the support of a few centers of study and exploration of these collective endeavors can be supported. One or two are already in operation.

Educational testing can be a useful aid to contemporary education or it can be an impediment. The choice seems obvious but the necessary programs of action must still be developed and implemented.

TESTING AND THE NEW TECHNOLOGY

Arthur S. Melmed
Office of Special Studies
National Institute of Education

For various reasons, historical and practical, the norm-referenced, multiple-choice, machine-scored paper-and-pencil test has been the dominant testing technology in use in the U.S. for many decades. Recently intense debate has developed over how effective and satisfactory an implementation we have achieved in the multiple uses to which we put this technology.

In considering alternatives to the use of the multiple-choice, paper-and-pencil test, cost must inevitably be a factor. The purpose of this note is to describe new information-handling devices, whose flexible, low-cost application to testing, teaching and learning lies within the planning horizon.

A number of papers in this volume discuss specific examples of the application of modern information-handling technology to educational testing. There follows some background on the information-handling technologies themselves, and other educational applications.

1. Technology

We have entered a revolutionary age in the technology of information handling. The microelectronic revolution, extensively documented in the September 1977 issue of Scientific American, is making possible a roughly tenfold decrease each five years in the cost of the integrated circuits that are at the heart of contemporary digital computers. We have available for individual use, at relatively low prices, sophisticated calculators and computers and means of generating images on video tubes. As supplements to the now universally available TV receiver, we have video tape recorders and players and will soon have the even cheaper, and in educationally-important ways more flexible, videodisc player. Following are some examples of currently available devices useful in education:

a. Computers and Calculators

Handheld calculators providing the four arithmetic functions plus square root now sell for about $10, and the cost of calculators providing trigonometric and logarithmic functions have declined to almost the same level. Calculators able to handle 400-step programs are now at the $100 level.

Personal computers, which include a TV tube display, keyboard, processor, and sizable memory, and weigh less than 50 pounds, are now available for as little as $600. Accessories include word-processing devices allowing flexible construction and editing of written text and production of typed copy. These computers are comparable in capability to computers costing hundreds of thousands of dollars a decade ago, and are only the first of many that will become available at decreasing prices or with much greater capability at the same price. Incorporation of microcomputers into TV sets could make substantial computer power available in the home to each child.

b. Video storage, retrieval, and display devices

A videodisc system coming onto the market for $700 provides the following, in conjunction with a regular TV set:

- 54,000 separate frames of full color image, alphanumeric or computer information on one side of a disc similar in size and cost to an LP record. The 54,000 frames correspond to 30 minutes of video at 30 frames per second.
- Rapid (dialed in more expensive models) random access to each of the frames, which are individually numbered and can be viewed individually for any length of time desired and may contain roughly one quarter of a page of easily legible text (or a full page with special high resolution video tubes).
- Easy replay, at regular or slower speed, of any portion of a motion picture.
- Limited programmable control of sequencing of frames with internal digital logic or expanded program control via a separate external microcomputer. Such combinations of varying complexities are referred to as intelligent videodisc systems. None is available commercially at this time, but the National Science Foundation has let two contracts for development of prototypes designed to be useful in education.

The disc can store 10 billion bits of information (Encyclopedia Britannica contains two billion bits; a human chromosome has a capacity of about 20 billion bits, and the human brain perhaps 10,000 billion). This immense storage capacity could probably include on a single disc all the computer courseware ever published.

There are now available, for less than $1000, home color TV tape recorder-players. Tape systems have the advantage of allowing recording and playback of non-pre-recorded video program materials, but tape is currently more expensive than a disc, and single frame display is more difficult in tape systems.

c. Other technology applicable to education includes:

- Handheld instructional devices providing drill and games in arithmetic ($15-25) and in spelling, with a simulated human voice ($65), are selling well.
- TV games, selling for as low as $40-50 are a familiar form of video graphics. Computer-controlled graphics with 15 colors are available for $300-400.

- A speech generator that strings together common phonemes in response to computer commands is available for under $400., and a device which can be taught to recognize a number of spoken words starts at $300.

- Music generators now range from ones which string together single notes into a melody to those which permit control of attack, decay, and timbre.

- Word-processing devices that allow preparation of hard copy of written text after editing and margin justification are available in various forms, including as an accessory to home computers.

2. Educational Uses and Implications

Existing and other foreseeable and barely foreseeable devices make possible great improvement and cost reduction in existing educational applications of technology, and also major qualitative changes in the use of technology.

Following are examples relating specifically to basic skills in reading and computation:

a. Drill and Practice

Currently available full computer-based drill and practice programs in elementary reading, language usage, and computation were designed to be economically feasible with the costly technology of the 1960's. At current or projected prices of computer capability, very considerably more sophisticated programs could be produced for delivery at the same cost. Alternatively, the same program should be deliverable more conveniently at a considerably lower cost (e.g., it seems feasible to build an entire computational drill and practice program into a hand-held calculator costing perhaps $50 or less).

b. Testing

Much standard testing is now carried out in the multiple-choice format to facilitate either hand scoring or machine scoring based on half-century old technology. Current and prospective technology makes possible computer-aided diagnostic and performance testing of far greater sophistication at costs comparable to tests based on the old technology.

c. Instructional Programs

The combinations of microcomputer and videodisc (with video and audio) makes possible sophisticated instructional programs at low cost, including instruction of students who cannot yet read instructions. These programs are generally designed to be useful to the teacher as a supplement to regular classroom education. They can be advanced examples of existing types of CAI programs, but new low cost computing power and new knowledge of human learning processes makes feasible qualitatively different sorts of interactive instruction in which the computer is "knowledgeable" about:

- o the subject matter being taught, and
- o tutorial strategies suitable for dealing with learning behavior of individual students.

It is easy to imagine computer-videodisc programs teaching reading comprehension strategies in ways that have not been possible in the past. Early work is underway on programs that allow children to learn mathematics (or reading and writing) in the same way they learn to understand and speak their native tongue - by "immersing" them in a world where mathematics is "spoken".

Speaking generally, these technologies substantially improve our ability to:

a. keep students actively engaged in learning for sustained periods - a prime requisite for effective education;

b. allow teaching to be adjusted to the rate and style appropriate to the individual student and to the particular difficulties the student encounters.

This brief review of modern information-handling technology and its uses and implications for education is intended to suggest what is possible in the way of assistance to the classroom teacher and the learner. Development and research are necessary to realize this promise.

II. EDUCATIONAL OBJECTIVES AND COGNITIVE MODELS

A DISCUSSION OF TESTS THAT CONCERN THE SCIENCES AND THE TECHNOLOGIES

Jerrold R. Zacharias
Founding Trustee, Education Development Center

Abstract

Although some of the processes and some of the content of the sciences and the technologies should be learned, or available for study at almost all student ages, no tests now available give a decent representation of them. New tests should be prepared. They should be part of a system which includes learning, teaching and testing, and they should be contrived to be fair, open with easy access, of high quality regardless of the purpose, the sophistication or the grade level intended. This can all be accomplished by making tests which are one third mandatory, one third partially mandatory and one third fully free. A research, design and development program could accomplish this miracle in short order (six years) and for a small budget ($6 million per year).

Science and technology -- like history, literature, and other studies -- can be thought of as tapestry subjects; that is, each entails a richly textured, interwoven, multi-dimensional and growing body of content that is widely varied. Isolated pieces of tapestry may appear to be very different from one another. In its entirety the tapestry is so large that it cannot be seen in very much of its minute detail at one time. It is tempting to point out that some parts of it are of almost universal interest and relevance, -- to every one of the students a knowledge of one's own body, how it is put together, how it works, how it got that way, how it interacts, etc., etc. seems just exactly appropriate. But it would not have been for me as a student, nor even now. The cosmos, surely everyone lives in the cosmos, -- but it is not even now claiming much of my attention. And interests come and go.

This aspect makes testing students for their familiarity with content, as opposed to process, an unfair unwieldy business.

A workshop held at the 1977 convention of the National Association of Elementary School Principals was devoted to mathematics, the technologies and the sciences. The participants were asked to focus on a number of questions -- and the recorded answers turn out to be relevant to thinking about tests in these subjects.[1]

Question 1: What is the role of subject matter in a science and technology curriculum? Is it an end in itself, or is it a vehicle to help people learn to think more critically?

---

[1] Judah L. Schwartz and Jerrold R. Zacharias, National Elementary Principal, 37, 1, 1977 October

The answers to this question, as summarized by the discussion leaders, are quoted in part here:

> Subject matter should not be an end in itself.
>
> Content should be considered secondary.
>
> In the beginning or primary grades, a science and technology curriculum should emphasize applications to the world around us. And there should be increased emphasis on the process of critical thinking to make use of the different facets of knowledge learned.
>
> Subject matter should be the vehicle and not the end. Sometimes teaching can get too hung up on content.

Clearly the elementary school principals in that workshop understood that neither teaching nor learning nor testing should focus on stipulated and required narrow subject matter no matter how tempting it may be to do so. But they wavered a little as is evident in:

Question 2: Should there be a national science and technology curriculum? How can we strike a reasonable balance between diversity and uniformity?

Most of the participants agreed that there should be no national curriculum. The discussion leaders summarized their comments as follows:

> Actual objectives should be decided by the local district.
>
> We do not need a national curriculum, but guidance in branching out to open teaching for flexibility via suggested concepts.
>
> A national program that stresses the inquiry approach seems desirable. Emphasis should be placed on teaching critical thinking skills and applying them to all disciplines.
>
> Yes, if it is not too specific, perhaps there should be a national curriculum. State-level goals would be better, however, so that there would be diversity. Regional science programs as well as national programs should be coordinated. Conclusion: something local in character but with national coordination, but no mandated national curriculum.
>
> The materials and tests now used constitute a basic program. There is a core of material the children across the country are already exposed to without a national curriculum.
>
> No single curriculum will serve all of the needs of every child. The time frame, social changes, and needs of children influence the curriculum.

There was consensus that there was no need and no desire for uniformity in the curriculum areas of science, mathematics, and technology. There seems to be almost complete agreement that there should be diversity in these curriculum areas, but that conclusion, too, requires several comments.

How many of something does it take to provide diversity? If you think of a smorgasbord, how many different types of cheese should there be in order to provide the best selection? Two? Ten? Fifty? A hundred? Common sense tells us that some single-digit number may be adequate -- at least, in terms of curriculum. But that number depends on how diverse is the taste of the children, the teachers, the principals, and the school system, and for whom does the diversity serve. Can every child or small group of children choose what they want to work on? I think that they should, with two provisos: 1. There should be no option to do nothing. 2. What they choose should not scare the teacher to death.

So we conclude that, for children in elementary school, at least, tests should in no way depend on familiarity with specific subject matter even if it is as familiar in the general culture as the germ theory of disease, or the atomic nature of matter. And a general familiarity with the standard chestnuts, frequently called science, should be either not used or given in sufficient profusion so that a student can reasonably expect what he or she will be asked. Procedures for making fair, diverse and first class arrangements for testing "general knowledge" are discussed a bit further along.

Can we turn to those features of the sciences and the technologies that differentiate them from say "humanistic studies" or music?

What makes the tapestry of the natural sciences an organic whole? Certainly not its surface features, which may vary widely. Many scientists believe that its richness and breadth and unity stem from common ways of observing, shared beliefs about the nature of evidence, and a set of intellectual tools and strategies that are widely used. For instance:

1. A basis for belief stems from observation and evidences of widely different sorts. How varied the sources of evidence determines how firm is the belief. I would wager my life against a dime on the correctness of an atomic theory of matter but I would give only even odds at small stakes on a theory of how snowflakes grow. Clearly decent respect for the various nature of evidence is a cohering pattern in the tapestry. Could we test for it? I think we could. With five-fold multiple choice test items, or even machine-scorable before 1980? I think not.

2. There are other ways in which study of or participation in the natural sciences or the technologies can help develop intellectual strengths, intellectual integrity and intellectual agility. The following is one recent form of a partial list of them:

- -- Is it really a problem that, in principle, can be solved?
- -- What can be ignored for the time being?
- -- Is there a quick, fast guess? And how can it be improved?
- -- Are there subproblems that can be solved first? (Does the problem break down into smaller problems?)

-- Pushing to the limit; pushing to absurdities
-- Establishing upper and lower bounds
-- Bringing in outside information
-- Recognizing and exploiting symmetries
-- Changing point of view and/or frame of reference to simplify the problem
-- Changing scale to gain a different perspective
-- Shifting attention betwen parts and wholes to develop new approaches
-- Anthropomorphism: What would I do if I were a...?
-- Taste, style, and elegance; seeking out clean approaches and solutions
-- Asking for help: Why, when, and of whom; understanding that it is no disgrace to ask a question, or to be unable to answer one.

Could we test for ability, agility and appropriateness in exploiting any of these ways to rummage around in the intellectual attic? I think we could, not with inappropriate mechanisms, of course, but somehow, and that how depends completely on carefully planned arrangements for a join of teaching, learning and testing. Each of these must be supportive of the other two. I want to leave to the reader the enjoyment of finding enjoyable examples for each of the ploys listed above. Successive approximations are the easiest, and most omnipresent in everyday life as well as in the sciences, the technologies or mathematics. Since this collection of ploys forms the basis for most intuition, which in turn requires substance (or entries) in the memory store, the working fluid of the test arrangement is heavily dependent on students' interests and sophistication. Nonetheless -- teaching -- learning and testing could be devised. Tests which are independent of teaching and learning, of general and blind application, could be a disaster.

In everything I have said so far, I have kept in mind an image of a ten-year-old child, boy or girl, capable of listening, speaking, reading and doing -- but with restricted ability to write. Suppose now we shift attention to students in secondary school; where the Second Tower of Babel controls; where courses of study are supposed to be unrelated to any other course; where there is indeed a nationwide uniform curriculum set by a half-dozen publishers and purveyors of tests; where no choice is allowed to any teacher, student or school; where science is regarded as available in smatterings and indeed only for the college-bound (no pun intended) student.

I believe that there is a way to break out of this box; that it could be done in a few years at a cost of no more than a half-dozen million dollars per year for a half-dozen years; that a concentrated three-month, sixty-person workshop could get it started within a half-dozen months of the word "go," flashed in green; that the transition from where we are now to a reasonable position could be gradual and non-disruptive; that the resulting tests could be fair to everybody; that they need not be secret; that they could (and should) result in non-numerical letter grades, preferably three of them standing for "super," "medium" and "poor." Finer tuning than this of test grades is a farce anyway.

Let's turn immediately to a concrete proposal and later put it in context by referring to the outline for discussion of achievement tests which appears at the end of this paper.

Begin the process by collecting, compiling, correcting and categorizing (several sorts of categories) a mountain of existing test questions with enough new ones to make the process one that is interesting enough so that it does not just fall into the hands of human robots without imagination or insight. Choosing the people who are in control will determine whether the questions are "serious" or just solemn. Three first class names come immediately to my mind, -- E.M. Purcell, George Miller and Richard Crane. So do names of people to avoid -- but since they are covered by the name "Legion" no reference is made.

So now assume that we have easily accessible (computers are wonderful) n questions available for $n!/(n-m)!$ tests with m questions. Full and easy public access not only to the test items themselves but also to typical collections of items that we call tests. We use this public question bank in three different ways, in order to minimize secrecy, and maximize diversity and choice on the part of the test-taker. One third of each test should be regarded as mandatory, a set of items with no choices available. Answer every question or else. One third of each test should be part mandatory, part permissive. For instance, -- "choose any five out of twenty." One third should be completely permissive. "choose any set of questions you wish." Pick them from any list or make them up to fit yourself.

Such a system could be self-purifying provided there is some mechanism to have the quality-derivatives go positive. If they depend on popularity ratings alone, then, to quote Andre Gide on seeing a performance of Cyrano, by Rostand, "Every generation gets the Shakespeare it deserves."

The various and interrelated issues that must be considered about this proposition are as presented in this outline for discussion of achievement tests:

I. A report about scope requires a multidimensional statement including:

- -- Elementary
- -- Junior High
- -- Secondary
- -- College Entrance
- -- Minimum Competency
- -- What is out there now? --short term
- -- Mathematics and the Natural Sciences
- -- Social Sciences
- -- Language (Listening, Speaking, Reading, and Writing about Something) vs. Copy Editing
- -- What is likely to be there in the future? -- near term
- -- What should be out there? -- near term, long term

II. General properties of tests
- -- Validity, item by item, vs. total test as a unit
- -- Secrecy vs. openness, with easy access
- -- Quality, both pedagogic and substantive

-- Diversity and choice vs. regimentation and requirement
-- Levels of sophistication
-- Class or ethnic bias
-- Talents biases

III. Purposes of tests
-- Diagnostics with fast feedback, or part of a teaching, learning and testing arrangement
-- Evaluation of projects, courses, systems, etc.
-- Accountability, keeping the quality up
-- Gatekeeping: a process rather than a test; the roles of the dossier and weighted lottery
-- Licensure
-- Reallocation of funds

IV. Costs: dollars, time, effects on children, teachers, etc. Can "good" tests be made, administered, and interpreted?
-- Who's in charge?
-- Who sets level and tone?
-- How disseminate the real sophistication about tests?

V. Relation between teaching, learning, and testing

VI. Access for students and relevant adults to tests as well as to scores -- legal questions

VII. The present thorns in the sides
-- SAT decline
-- Numerical grading: the one-dimensional array nonsense
-- IQ and other mental agility tests
-- Multiple choice

VIII. The out-of-school tests: CBS Science Quiz show, magazines, press

IX. The special case of the "tapestry" subjects

X. Criteria for criterion reference, etc.
-- What, who, and how?

XI. Should the government help finance a broad diversity of tests? Who, how, when, where, etc.?

XII. Computer and other technologies (videodisc) applied to testing

XIII. What we should not be tending to:
-- IQ or mental ability tests in any form
-- Gatekeeping? Role of lottery
-- Bakke issues
-- International tests? Thorsten Husen
-- Accountability
-- Role of Ed. Psych.

XIV. National Consortium on Testing

ASPECTS OF MATHEMATICS LEARNING THAT SHOULD BE THE SUBJECT OF TESTING*

Robert B. Davis
Curriculum Laboratory
University of Illinois, Urbana

with the cooperation of

Leon Henkin
Department of Mathematics
University of California, Berkeley

The title suggests, correctly, our belief that there is more to pre-college mathematics than is commonly dealt with in examinations. Neither the range of behaviors, nor the range of topics, is usually adequately explored.

I. METHOD

Our fundamental method is to assume that an interested mathematician, were he able to spend three or four days with a student (and if he had experience in such matters), could arrive at a reasonable assessment of what that student knows, understands, and can do in the area of pre-college mathematics. During those three or four days our hypothetical mathematician would, presumably, do a variety of things: ask the student to write out solutions to certain problems, or to write out proofs of certain theorems; sit with the student and go over some of the student's written work, asking questions in order to probe more deeply; possibly ask the student "essay" questions, such as: "What should students learn about quadratic equations, and why?"

In this note we try to describe, at least approximately, what we believe our hypothetical examiner is trying to find out.

It should be mentioned that there are inevitable implicit assumptions about test construction, test administration, the reporting of test outcomes, and the uses made of such reports that influence any discussion of testing. Some of these considerations are mentioned in Appendix A. We believe, however, that the main core of our remarks are relevant to any reasonable set of such assumptions.

Even more importantly, we try to discuss different kinds of thinking that are part of mathematics. This is treacherous ground, for many reasons. The reader may wish to challenge our analysis of "mathematical thinking" Here too, however, we believe that the main kernel of truth in our remarks would remain true even in the face of a virtually total rejection of our tentative analysis of mathematical thought.

The key question, then, is not the literal correctness of our entire analysis, but rather whether we are pointing to appropriate and important kinds of test items. Are we, in fact, proposing to test the student's learning of mathematics in some appropriate sense?

---

*The relationship between this paper and the next is described on page 84.

Primarily for clarity of communication, we discuss half a dozen (or so) aspects of mathematical thinking, which we identify as:

knowledge of facts
algorithms (or VMS sequences)
concepts
frames (in Minsky's sense)
meta-discussions of algorithms
heuristics
linguistic aspects of mathematics

and something we have come to refer to as "larger systems."

We also focus considerable attention on the notion of "understanding." Although we are fully aware of the tendency in contemporary education to dismiss "understanding" as hopelessly mystical, we believe that our hypothetical mathematician-examiner will want very much to know whether the student "understands"--and we are confident that this is not meaningless, but on the contrary can be subjected to reasonable explication.

II. THE VARIETIES OF MATHEMATICAL THINKING

A. Facts.

The most rudimentary kind of thinking deals with facts: a first grader, seeing 3 + 4 =, learns to respond by saying or writing "7". For the first grader this may be a complex cognitive process, as the work of Suppes and Groen suggests--but for the more advanced student, dealing with quadratic equations, the response "7" is so familiar that it is largely automatic. In the present discussion we treat such responses as[1] "unit" processes that we do not attempt to decompose analytically, except in those cases where further probing may be required.

For testing purposes, facts are relatively untroublesome. But, as Hoffman, Hawkins, and others have pointed out, even quite "simple" facts tend to embody a surprisingly large number of assumptions. Hoffman (1962) gives the example: "True or false: George Washington was born in 1732."

Prior to careful scrutiny that information might seem to be entirely factual, carrying no significant burden of assumptions. A little consideration shows a quite different picture: true, *if* you count from the present (1978) backwards, you conclude that George Washington was, indeed born in 1732. However, between the time of Washington's birth and the present, a calendar reform occurred. If, say, you are making a movie about the birth of George Washington, the calendar on the wall should NOT say 1732.

---

[1] The categories of mathematical thought used here have been influenced by the analysis presented in Davis, Jockusch, and McKnight (1978).

This is NOT an entirely trivial complication. If a student says that the equation

$$x^2 + 1 = 0$$

has 2 real solutions, we suspect our inquiring mathematician would want to know whether the student misread the problem as $x^2 - 1 = 0$ (or has made some other "accidental" error), or whether he is ignorant of the meaning of the word real (a term involving a classification of numbers as real, imaginary, or complex), or whether he has never noticed that some quadratic equations do NOT have 2 real roots. In this case, our mathematician may want to see if the student can conclude anything from, say, a thoughtful examination of the quadratic formula.

We suspect that in many cases a "factual" question would tell us more about the student's mathematical thinking if it could be followed up by further probing of the type we have just suggested. If a test is administered by a computer, this follow-up probing becomes a definite possibility.

B. Algorithms. [Our main discussion of algorithms appears in the section written by Henkin, but the subject is so important that it is also discussed briefly here.]

Perhaps the most prominent kind of process is the algorithm or visually-moderated sequence ("VMS").[2] Long division is a typical example:

The notation

$$17\,\overline{)\,1978}$$

triggers an automatic process of asking "how many times does 17 go into 19" (or something to that effect), which leads to a revised notation

$$\begin{array}{r} 1 \\ 17\,\overline{)\,1978} \end{array}$$

but now this revised notation, which had been the output of an earlier process, becomes the visual stimulus or input for the next cycle, during which one multiplies 1 by 17 and writes

$$\begin{array}{r} 1 \\ 17\,\overline{)\,1978} \\ \underline{17} \end{array}$$

---

[2] Davis, Jockusch, McKnight (1978).

But now this new notation becomes the visual trigger for the next cycle (namely, subtraction of 17 from 19, yielding the result 2), and so the process continues, until some visual stimulus calls a halt.

This kind of algorithm or VMS sequence appears, in general, to pose no great problem for the usual kinds of analysis of S-R bonds. It seems to be the major ingredient in pre-college instruction, perhaps to a harmful extent; it is also something that many students come to expect. Indeed, there is evidence that when mathematics instruction at the college level (or elsewhere) begins to include processes other than this, some students are quite uncomfortable that their firm expectations are being challenged.

Testing for algorithms is perhaps the least complicated of all mathematical testing: one asks the student to solve a problem which can be solved by the algorithm in question. Even here, though, there is some small room for confusion. Most problems can be solved by more than one algorithm. If some specific algorithm is the goal of the testing, some effort must go into selecting the problem to be solved--or some additional instructions or further probing may be required. It is not merely the ability to obtain a correct answer that is required--some knowledge of a student's repertoire of algorithms is important, particularly if the testing is to serve some diagnostic purpose in describing either the needs of this individual student, or else the strengths and weaknesses of the curriculum. As an example, a student might be able to find 50%, 25%, 20%, and 10% of, say, 500 by using an algorithm which would NOT allow him to translate "X% of N" into correct algorithm notation. It is important to have accurate information about a student's actual repertoire of algorithms.

There is another complication that might be mentioned here: we know a number of instances where a student has concluded that he or she cannot solve a certain problem, and has ceased working on it. Given a few well chosen words of encouragement (with absolutely no mathematical hint in them), the student has resumed work and solved the problem successfully. For any form of testing this is a very important phenomenon; for diagnostic testing it is critical. Such students might appear deficient in knowledge, whereas in fact what they lack is self-confidence or determination.

There is another frequently-observed behavior: on tests with student-constructed answers, if a test proctor merely indicates which problems are wrong, many students can then go back, identify the error, and correct it. [This is one more case where administering exams by computer could help.]

C. Concepts. There are, however, bits of mathematical thinking that do not appear to be algorithmic in the sense of the preceding section. The distinction would seem to be important.

Before attempting to define what is meant by a mathematical "concept," we give two examples, and ask what they have in common.

1. The "zero product principle"[3] states that, if the product of two numbers is zero, then at least one of the numbers must be zero. This concept justifies the step:

line (n) $(x-2)(x-3)=0$
line (n + 1) Therefore, either
$x-2=0$
or else
$x-3=0.$

[To be sure, at a more mature stage this step can be justified in terms of a proper use of variables in the statement of axioms and rules of inference

$$x \bullet y = 0 \rightarrow (x = 0) \vee (y = 0),$$

but in typical situations beginning students do not deal with the matter in this way.]

This notion of "concept" has interesting properties, and raises interesting questions:

a. Beginning students, taught with considerable emphasis on this concept as a justification for the algorithm that solves quadratic equations by factoring will, in a sizeable number of cases, learn the algorithm but will not be able to explain it in terms of the concept--despite the heavy emphasis that was placed on the concept during the instruction. These students display a high selectivity: algorithms (or VMS sequences) they will learn, but "concepts" they ignore. [It sometimes seems almost like color blindness in its specific selectivity.]

b. Experienced mathematicians, on the other hand, appear to give priority to the concepts, and believe that, by making use of the concepts, they can reconstruct the algorithm (and hence are not completely dependent upon memorizing each detail of the algorithm). How does the concept allow a re-creation of the algorithm? There may be an answer that can be stated in terms of Minsky's frames, but we do not pursue the results further at this point, since it would probaby lead us too far astray.

2. The concept of function. A mathematician has certain quite definite ideas that come to mind in response to the word "function." There is the present-day formal definition as a kind of set of ordered pairs; there may be recollection of earlier definitions in terms of "the value of y is known when the value of x is given," and

---

[3] This name for this important concept is taken from Stein and Crabill (1972); there is no name in truly common usage--one name, sometimes used, is "the non-existence of divisors of zero".

definitions that allow "many-valued functions" and "single-valued functions"; there are metaphors in terms of machines, such as the pictured hopper-and-spigot gadget in Begle's Calculus; descriptions in terms of "inputs" and "outputs"; common notation in terms of "x" and "y"; the notation f(x); Bourbaki's "allowable abuse of language" in the notation f(x); representations of functions by formulas, graphs, and tables; classifications of functions and methods for achieving these classifications; the surprisingly deterministic "rigidity" of analytic complex-valued functions vs. the "flexibility" of real functions; representations as infinite sums or infinite products; and very much more. The actual collection of ideas will presumably be different for different individuals, and is certainly very different for a mathematician and for a beginning student.

While we would not expect the immediate assent of a majority of mathematicians, some (at least) of the present authors find it attractive to think of the mental representation within the mind of one person of function--the "idea" of function--as one of Minsky's "frames", a relatively new concept that merits some discussion.

Frames [sometimes alternatively called "schemas" (Rumelhart and Ortony (1977)] or "scripts" (Abelson, 1973; Schank, 1975)] were proposed by Minsky and others [Minsky (1975)][4], partly in the context of artificial intelligence and partly in the context of cognitive psychology, to deal with several troublesome problems, including these:

a. The "combinatorial explosion": if information is stored in memory in the form of a very large number of very minute fragments, the searching process can involve so vast a number of decisions as to become entirely unmanageable;

b. Consider what happens when a typical U.S. adult reads a specimen passage such as this:

> It was Paul's birthday. Jane and Alex went to get presents. "Oh, look!" Jane said, "I can get him a kite." "No, don't," Alex said, "he already has one. You'll only have to take it back."

The typical adult reader can easily answer questions about this passage, such as:

Q_1: Where did Jane and Alex go?

Ans: To a store that sells kites and other things.

Q_2: Why are they buying presents?

Ans: As birthday presents for Paul.

---

[4] See also Winograd, 1977.

Q_3: To what does the last "it" [the next to last word in the passage] refer?

Ans: A new kite, if Jane bought one.

This typical reader is quite surprised when he is told that neither of the first two questions are in fact answered in the specimen passage. [It is also by no means easy to devise a linguistic rule to answer the third question.]

Where does this extra information come from? Minsky's answer is that something, perhaps the first sentence, has triggered the recall of a rather larger, internally-structured, aggregate of memory-stored information. This aggregate might be called the "birthday frame". This internally-structured "birthday frame" provides for "whose birthday?", "when?", "where?", "how old?", "what presents?", and so on, in the form of variables.

What is involved, then, is:

(i') selection and retrieval of a possibly appropriate frame;
(ii') replacing some of the variables in that frame with values determined by this specific information input;
(iii') making some sort of check for the suitability of the frame selection and the variable replacements;
(iv') thereafter basing further information processing on the frame (including the specific values assigned to the variables).

A default procedure assigns values to variables, when necessary, if the present information input fails to do so.

Hence, if a mathematician retrieves a "function" frame, that frame immediately suggests questions such as "How represented?" "What kind of functions?", and so on.

We are now in a position to answer what the "zero product principle" and the concept of "function" have in common: our answer is that each seems to play a frame-like role.

This requires some explanation.

To begin with, it is clear that neither "function" nor "zero product principle" have the form of a tyical VMS algorithm; in neither case do you start with a visual input and respond by writing something. [Note that you would probably be willing to explain either "function" or "zero product principle" without pencil and paper, but would probably not be willing to explain long division in a setting where you could not write.]

The fastest road to clarification may be to watch this "frame-like" role in action. Suppose a student, S, is presented with

$$(x - 2)(x - 3) = 0.$$

Then, if S has learned what we hope, S will think of (or "retrieve") a "zero-product" frame, probably involving something like

$$(\underset{\text{something}}{\qquad\qquad}) \times [\underset{\substack{\text{something}\\\text{else}}}{\qquad\qquad}] = 0$$

This frame carries the further information that one or the other of those factors must be a zero. S matches up the variables with his specific instance, finds the match acceptable, and now makes use of the information contained in the frame.

In the case of "function", we have already seen some of the questions that the "function" frame may ask us to answer from the available input information.

An important further property of frames is that they refer us to a relatively small number of other frames, rather like the "see also..." references in an encyclopedia.

What test items would you use in order to determine whether or not student X "has the concept of...", or "is familiar with the idea of...", some topic Y?

This is a question related to, but different from, testing for algorithms. Frames presumably direct us to a sometimes sizeable collection of specific algorithms, and some of these might be worth testing for--a matter we have discussed earlier. But you perhaps want to test also for the concept, which is NOT identical with testing for algorithms. Actual observation of students has revealed many students who can solve some quadratic equations, but cannot answer questions such as:

> Give an example of a quadratic equation.
> Where, in the study of quadratic equations, do we make use of the zero product principle?
> In the expression
>
> $$ax^2 + bx + c = 0$$
>
> in what way are a, b, and c different from x?

Observations suggest that, at the present time,

- pre-college students seem to expect mathematical activity to consist mainly (or entirely!) of using algorithms

-- pre-college tests test mainly for the ability to use algorithms

-- pre-college instruction places a heavy (and sometimes nearly exclusive) priority on algorithms

This has the appearance of being harmful to students, because it leaves them unprepared for university-level mathematics instruction.

It consequently appears that testing should attempt to increase the amount of emphasis given to testing for concepts. This should NOT take the form of items that test primarily vocabulary. Probably our hypothetical mathematician-examiner would ask some rather open-ended questions, such as:

Q1: What methods do you know for solving quadratic equations? When do you use each one? [Or: how do you decide which one to use, for some given problem?]

Q2: If I know a vector $\underline{v}$, and I want to find the component of $\underline{v}$ in the direction of a known vector $\underline{w}$, how can I do it?

Q3: How could I find the distance from the parabola $y = x^2$ to the point (8, 1/2)? [For most students this would not be a "type problem" which they could solve by the application of known algorithms.]

Q4: What is the logical justification for each step in the following calculation?

Or, conceivably, even questions such as:

Q5: If you were going to teach someone about quadratic equations, which specific things would you consider most important for him or her to learn?

Q6: Sketch the main sequence of theorems that together lead to the theorem of Pythagoras.

Answer [following Moise, or Anderson, Garon, and Gremillion, 1973]: The proof depends upon some algebra and on the area of right triangles. The area of a right triangle is found from the area of a rectangle. The area of a rectangular region is found by a diagram that relates it to the area of a square, which is proclaimed as a postulate.

We are aware that questions of this latter type are sensitive to the particular curriculum that a student has pursued. [Many other types of items have this curriculum-sensitivity, as well.] Since we consider it important that nation-wide testing should not discrimiate against responsible innovations in curricula, and since even differences among present curricula pose testing problems, we recommend a suggestion of J. Robert Cleary, of Educational Testing Service, that major curriculum alternatives (including new curricula in developmental stages) be listed, and test items be coded so that a student can be placed in relation to the curricula he or she has studied, with test items being selected for compatibility with those particular curricula. Tailoring item-selection to the individual student is another situation where computer-administered tests could prove helpful.

D. Heuristics

Every step in a mathematical proof (or in most kinds of calculations) must be justifiable by some rule of logic which, in effect, tells us what is, or is not, legal. But at a typical step there will be a quite large number of legal next steps. Logic says nothing about which of these next steps would be helpful in moving us toward our goal. What does attempt to offer such suggestions are principles of a different sort, known as heuristics.

A heuristic is less a rule than a suggestion. It may or may not prove helpful in any specific instance.

Examples:

1. In the derivation of the quadratic formula, one often includes the steps

$$ax^2 + bx + c = 0 \qquad (2)$$

$$x^2 + \frac{b}{a} x + \frac{c}{a} = 0 \qquad (3)$$

Why do we choose to transform equation (2) into equation (3)? [Answer: Because for many purposes, including completing the square, the task is easier if the coefficient of x^2 is 1.]

2. Some typical heuristics:

Try to decompose the problem into component parts, and try to solve each part separately.

Think of an easier, but similar, problem and see if careful consideration of the easier problem will help you find a way of solving the original problem.

Try to identify as precisely as possible that specific part of the problem that makes it difficult, then try to deal with this difficulty (for example, perhaps by eliminating it). Try to think of a similar problem that you already know how to solve, and see if you can transfer this new problem into the familiar one.

> Imagine that you are in the middle of the process and try to figure out how to proceed from there--this often works better than thinking first about the actual beginning of the process.
>
> What is conspicuous by its absence? Can it be introduced into the problem?

There are also many "almost" principles, of a kind suggested by Littlewood, such as: Every analytic function is almost a polynomial. Every continuous function is locally almost a constant.

Far too little attention has been paid, both in testing and in teaching, to the matter of heuristics.

We have included above one example of a test item dealing with heuristics ("Why did we transform equation (2) into equation (3)?") That item presents a decision and asks for a heuristic justification for the decision.

Alternatively, one can present a heuristic and ask for an instance of it, as in:

> In answering the question "What is the smallest value of y?", given the equation $3y + 5 = 8x^2 - 10x$, how could you use the heuristic device of "turning it into an easier (or more familiar) problem"?

Heuristics are not, at present, widely taught. How, then, should testing approach them? If testing continues to ignore heuristics, so too will instructional programs--despite the great importance of heuristics in achieving a genuine command of mathematics. Yet one cannot fairly test that which is not taught.

Undoubtedly some transitional aproach can be developed: for a period of time, emphasize the kinds of heuristics questions that will become a part of tests. Then for a period of time, include a few such items, but only as "extra credit," not as part of the obligatory minimum core of the test. Finally, include them as regular test items.

E. *"Larger Systems"*. [The main discussion of this topic appear in the section written by Henkin, but it is so fascinating a topic that one cannot resist discussing it briefly.] We have become fascinated by how often, in the discussion of a specific test item, we have found it making contact with what, for want of a better term, we have come to call "larger systems". Perhaps the broadest [and therefore not-quite-typical] "large system" is *reading and writing*. If you cannot read or write, you cannot deal satisfactorily with test items. A more typical example would be the classification of statements as *true* or *false*. If you are not adequately familiar with the classification of statements as "true" or "false", then you cannot truly understand mathematics, no matter how carefully you reproduce the hieroglyphics of mathematical symbolism.

To ask if someone understands the classification of statements as "true" of "false" may seem foolish--it can appear that surely everyone knows this. In fact, however, not everyone does. There are many different ways of classifying statements, the mathematician's "true" vs. "false" being only one of them. Many people classify responses as "right" or "wrong", and then attempt to define true and false in terms of their notions of right and wrong. This usually fails, because the common notions of right and wrong have no simple translation into true and false.

Another of these "larger systems" involves the matter of equations. If you do not understand the conditional nature of an equation, you can hardly be said to have "solved" an equation, even though you may have written

$$x = 2,$$

thereby correctly indicating the root.

There is the matter of variables: if you do not have an adequate understanding of how "variables" work, including the "rule for substitution" (as it is sometimes called), then you can hardly be said to have "solved an equation". What you have done, at best, is to write something that you do not understand. Observation of students indicates that "writing something you do not understand" occurs very frequently. This is suggested, for example, when a student writes

$$(x - 2)(x - 3) = 0$$

$$x = 2, 3$$

and is unable (as many are) to explain what he means by "$x = 2, 3$".

But a quadratic equation is a special kind of polynomial equation, and polynomial equations are a sub-class of equations. If a student does not know something about this system for classifying equations, can he truly be said to "understand quadratic equations"? The point is arguable; but an even clearer case arises if the student does not know how a, b, and c differ from x in the usual equation

$$ax^2 + bx + c = 0 \qquad (1)$$

or if he does not know that equation (1) stands for all quadratic equations, or if he does not know how to relate

$$x^2 - 5x + 6 = 0$$

to equation (1) by setting $a = 1$, $b = -5$, and $c = 6$ (and a surprisingly large percent of students do not), then he has NOT mastered "quadratic equations"--but the failing is not within the topic of quadratic equations itself, but rather in the "larger system" with which the topic of "quadratic equations" makes essential contact. A large percent of student failures in our observations are the apparent result of inadequate mastery of these "larger systems"--classifying equations, using variables and parameters, and so on.

F. Diagnostic Tests. If we want tests to tell us what a student does correctly, and what he does not do correctly, one thing we need is, in the sense of vector geometry, to look for those metaphoric "basis vectors" in the "space of mathematical skills" which are faulty. One or two faulty "basis vectors" can result in an inability to answer many questions on a test. In such a case, though many test questions be answered incorrectly, perhaps the student has only one or two "bugs" in his information processing. Remediation should then probably focus on just these few bugs, and the prognosis is probably favorable.

Considerations of this sort are a corollary to the discussion in Section E, immediately preceding.

G. Meta-discussion of Algorithms. Devising more elaborate algorithms by concentrating or modifying previously-existing algorithms requires some knowledge of the attributes of the algorithms themselves. Consider, for example, how you would find the distance from the surface

$$ax^2 + by^2 + cz^2 = 1$$

to the plane

$$rx + sy + tz = 1.$$

In order to select and properly sequence the appropriate algorithmic procedures, you must think in terms of descriptions of what these algorithms accomplish, and what they require as inputs.

H. "Understanding." We have used the word "understanding", which is sometimes considered unacceptably mysterious, if not mystical. Yet within mathematics, "understanding" is an unavoidable topic. If you do not "understand" what you are doing, then sooner or later that lack of understanding will reveal itself, perhaps in severely detrimental ways.

What does it mean to "understand", say, quadratic equations?

If you understand, you have a reasonably good knowledge of when each method can be applied. You can talk sensibly about the methods, and you can relate them to a lengthy list of other ideas: the derivation of the quadratic formula, the classification of equations, the logical justifications for steps, the heuristic reasoning behind each step, the relation of quadratic equations to

$$e^{2t} - 5e^t + 6 = 0$$

and so on.

But in talking with students, one of the most important distinctions that one becomes aware of is the distinction between the student who can laboriously work his way, one step at a time, through a procedure (possibly with some errors), vs. the student who seems to have the whole procedure laid out in his mind, who is able to discuss it, to explain it (both in logical and in heuristic senses), to compare or contrast it with other algorithms, perhaps even to modify it if necessary.

The distinction is sometimes described by saying that the first student has "verb knowledge" of the procedure--he can do it, more or less correctly--whereas the second student has "noun" knowledge of the procedure--for him it is a known thing that he can think about, describe, talk about, and so on.

Clearly this description makes "understanding" sound very much like possessing and retrieving an appropriate frame--when dealing with a frame, one can move around at will, whereas in dealing with an algorithm one can only move on to the next step.

Whether this interpretation in terms of frames is justified, it is certainly a fact that one has not mastered a mathematical topic until one has achieved the ability to talk about the relevant algorithms, to think about them, and not merely to carry them out one step at a time. This distinction is familiar to everyone, perhaps from trying to assemble some apparatus by following, one step at a time, the written instructions that came in the box.

III. SOME OTHER DIMENSIONS OF PERFORMANCE

The more commonly-used test questions appear to assume that one either can produce a certain behavior, or else one cannot. They seem to spend little effort in studying under what conditions can one produce this behavior.

From observation of pre-college students we have noticed that:

1. A student may be able to construct a correct solution to a problem if errors in his first (incorrect) solution are called to his attention. Of this student one might say: he cannot produce a correct solution in the absence of error-locating feedback, but, given such feedback, he can produce a correct solution. [This distinction is meaningful mainly where the number of possible student responses is so large (often infinite) that knowledge that an error exists does not trivially determine the correct respone (as it would, say, on a true-or-false question).]

2. A student may be able to construct a correct solution to a problem if he is given encouragement ("Oh, come on, you know how to do that! Think about it!") in a form that does not contain mathematical assistance.

3. A student may be able to construct a correct solution to a problem if he is adequately motivated to do so, but may fail to create a correct solution under typical levels of motivation.

4. A student may be able to construct a correct solution to a problem if he is allowed to refer to some written notes (as in looking up the quadratic formula, looking over a list of postulates in geometry, or looking up the definitions of the trig functions). This does not represent an unimportant kind of knowledge, since in adult life such tables can be consulted if necessary.

5. A student may be able to construct a correct solution to a problem if it appears in a context that is familiar and also uncluttered, but may be unable to do so if the problem is embedded in a context that is cognitively complex. [This would appear to represent a break-down of solution creation under conditions of high cognitive "overloading" stress.]

IV. SUMMARY

Mathematical behaviors that one wishes to test have an obvious content dimension: one might want to test long division, or solving quadratic equations, or performing algebraic operations on complex numbers, or estimating the number of fish in a lake.

But there are also two other kinds of dimensions.

The first of these deals with the kind of behavior that is required: knowledge of facts, knowledge of some specific algorithms (and the ability to use it), knowledge of concepts (which may resemble "possession of an appropriate 'frame'" in Minsky's sense), knowledge of heuristics (and the ability to use them), mastery of certain "larger" (or more general) bodies of knowledge, and understanding (which may also be explainable in terms of the ability to access an appropriate 'frame').

The remaining dimension deals with the question: under what conditions can a student solve this problem?

In general, there is important knowledge about a student's performance that can only be obtained by observing him as he constructs solutions to mathematical problems. Some of this may require "clinical observation" in the sense of Piaget and Ginsburg; some of it can perhaps be carried out by using computers to administer tests.

Appendix A

Assumptions Underlying Tests

A surprisingly large number of assumptions must be made in order to design a test, to report the test outcome, or to make use of this report. We present here a partial list.

A. Goal of the Test

Is the test a "certification" test, as (for example) the driving test is--to determine whether the candidate possesses some (possibly complex) ability, such as the ability to drive a car reasonably safely? Such tests are in a sense criterion-referenced tests, intended to provide a guarantee of some level of competence in some specific area.

Or, is the test intended to provide a description of the candidate's knowledge (or abilities)? Judgments made by baseball team managers are often of this type:

> Player X is a good hitter against left-handed pitching, but probably only for a one-base hit; but he is not a good base runner, so if he gets on, put in a runner; don't ask X to bunt--he gets better results swinging.

Or, as a third choice, is the test intended to locate a student's level of performance in relating to other students? Competitive admissions tests often have this character.

B. Underlying Assumptions about the Organization of Knowledge

Influenced by Piaget, some people assume that progress through some body of knowledge or capabilities occurs in a largely invariant sequential order: while different individuals progress at different speeds, the sequential order of skill or knowledge acquisition is exactly the same for everyone.

In such a case, one need only observe certain "check points". It is as if we were trying to locate a train along a one-track stretch: if it has passed check-point N, and not yet passed check-point N + 1, then it is located somewhere between N and N + 1.

The assumption of invariant sequential order may seem extreme, but assumptions somewhat like it are not uncommon.

C. Underlying Assumptions about the Curriculum

An assumption similar to the preceding one is sometimes made, not about knowledge, nor about human cognitive growth, but rather about the curriculum. It is often assumed to follow an invariant sequential order. Something of the sort is done in typical elementary mathematics tests, which report grade equivalents such as 10^{++} ["he does arithmetic about as skillfully as a typical 10th grader"], or 4.3, or 9.6, etc.

D. Assumptions about Learning and Forgetting

The monotonicity implied under C, above, involves also some assumptions about learning, forgetting, and practice. One would not be likely to assume this in the case of, say, U.S. history. Clearly, many students knew more about U.S. history at the end of the course than they knew one year earlier, or three years later.

E. Assumptions about the Representation of Test Outcomes

It is sometimes assumed that test outcomes can be reported by a one-dimensional linear reporting system. But a contrary assumption is at least as well justified, for example by the work of Erlwanger [Erlwanger 1973; 1974; 1975]. One expects a garage to report on one's automobile in terms of exactly what's wrong, and what it will take to repair it. This can be a metaphor for reporting on a student's mathematical performance.

It can also be the specific pattern of student errors that yields the most valuable information.

F. The Purpose of Reporting Test Outcomes

We may be seeking diagnostic information: what, precisely, does this student need, right now?

Or we may be dealing with a minor credentialling step, such as moving ahead to the next unit of instruction.

Or perhaps a major credentialling step, such as licensing a physician or an airline pilot.

We may want to give to the student some feedback on how well he is progressing in relation to other students. (Presumably this is a relatively undemanding evaluation task--typically one shows the student his grades on recent tests, or his standing within the class. The goal is merely to define, somewhat more clearly, what is expected of a student in this course or school.)

Appendix B

When Can Computers Help?

The use of a computer in testing would be especially valuable in certain situations:

1. When, as Zacharias says, you want to "go public" with a test: that is, make it available for public scrutiny, and still be able to use it. The computer can modify numbers. If desired, the computer can keep track of the difficulty of different number combinations, and can select combinations of about the same difficulty. [Alternatively, the computer can double as a calculator, and can help with the arithmetic, making all number combinations equally difficult.]

Besides changing details within a problem, the computer could have a very large pool of problems, and select a stratified sample. The large pool could be made public; the sample would differ every time someone took the test.

2. Keeping track of response time. I personally am rarely if ever interested in speed, but: whereas timed tests time rather sizeable pieces of work, a computer can time whatever tiny pieces of work one desires, in whatever combinations one desires. E.g., one could allow an unlimited amount of time for the student to read a word problem, and to set it up, and then one could measure the time the student takes to solve the equations; or alternatively, one could limit the time available for solving the equations.

The point is that the computer can measure precisely whatever portions of a test one wishes to measure, or can limit the time available for various specific parts of a test.

3. Alternatively, the computer could be programmed to allocate time in any way desired. This could be used to make sure that a student does not spend too much time on one part of a test. This need not be coercive--the computer could double as a watch, and advise the student (by flashing a message on the display screen) whenever he is taking too long on one part of the test--still giving the student the option of moving on, or not.

4. Of course, the computer could reduce the pressure of time, by allowing interrrupted sessions. A record of problems dealt with would be kept in the computer memory and whenever the student signed on again--perhaps at a different terminal, located somewhere else, but connected to the same computer--whatever further work the student did would be added to his record. When the student had completed the test, the complete results could be reported to the teacher or examiner.

5. The computer can also save time, by using responses to some questions to determine which later questions will be presented. Thus, if a student does not know that

$$(a + b)^2$$

is equal to

$$a^2 + 2ab + b^2,$$

there is presumably no point in even presenting a question about

$$(a + b)^3$$

or about

$$(a - b)^2$$

6. The computer can go a long way toward distinguishing between "careless" incidental slips, as against actual ignorance. For example, the computer can say: "Are you sure you added those numbers correctly?" or "Are you sure you read the problem correctly?", thus calling attention to what are presumably "careless" errors, that may not represent ignorance. The student would be given a chance to correct a wrong addition, or a miscopying of the problem, etc.

Such "help" can be as pointed, or else as general, as one wishes. The computer could point precisely: "You have an error in line 3," or it can point vaguely: "Your answer to one of the first five questions is wrong."

7. Lessening the reading problem--Method I. If the terminal has a touch panel, and an audio unit, it can read a sentence or phrase that the student touches.

8. Lessening the reading problem--Method II. The computer can rephrase a question--state it in different words, or explain it further--if the student requests it to do so.

9. Lessening the reading problem--Method III. The computer can display small portions of text, then more, then more, as the student requests it. The student is not then faced at the outset with a whole page of text--the text builds up gradually, as the student presses "Next." This may be less overwhelming for some students.

10. Lessening the reading problem--Method IV. In addition to suggestion 9, some text may lend itself to simple animation. "A test-tube is 1/2 full" can appear on the screen, and at the same time a picture of a test-tube can also appear on the screen, either half-filled, or becoming half-filled as the student watches. As more text is displayed on the screen, the animation can continue to show what is happening.

11. The computer can match the test to the specific curriculum. E.g., some books use the notation $^{+}3$ for "positive three," some use the notation +3, and some use 3. Some books do NOT consider that "3" is the same as "+3", but other books do.

This can make the construction of fair paper-and-pencil tests quite difficult. But a computer can deal with this easily, perhaps by asking the student, early in the test, what book he has studied.

Even more important, the computer can test the topics that student has studied. This is important. Today's paper-and-pencil tests often discriminate against various better curricula because the test doesn't match the curriculum very well. The computer can achieve a match.

12. The computer can match the test to student preferences.

In many cases, it doesn't matter which, of several topics, the student has studied, just so long as he has successfully studied one of them. Studying the biography of some U.S. president (or of some eminent scientist) might be a case in point. Perhaps it doesn't matter whether the student chose to study George Washington, or Abraham Lincoln, or F.D.R., or John Kennedy--the computer could question the student about whichever president the student selects.

13. Where sequential order of presentation is important.

Suppose in question 1 you want the student to fill in a truth table, then in question 2 you want the student to use the truth table to check the validity of some inference schemes. Now, of course, you want the validity check to use a correct truth table. With a computer-administered test, this is easy: after the student's response to question 1 has been analyzed and recorded in the computer's memory, the computer can present a correct truth table for use in question 2.

14. Where the sequential order of responses is important.

In some exams on medical diagnosis, the computer can present a patient's initial symptoms. The student can ask questions of the patient, or can order lab tests. When some test results are reported, further tests can be ordered.

The sequential order in which the student asks questions is the best measure of his performance--it shows how well he makes use of the information obtained from his questions and from the lab tests.

15. Simulation.

Obviously, there is almost no limit to the computer's ability to simulate various situations--driving an automobile, flying an airplane, etc.

16. Presentation of visual stimuli.

A computer can present nearly any visual stimulus material that is desired, via photos, animation, etc.

17. Presentation of auditory stimuli.

A computer can present nearly any auditory stimulus desired--in one case, for example, the stethoscopic sound of canine heartbeat.

18. The computer can focus on whatever part or aspect of student performance that is desired. For example, one can test student ability to interpret arithmetic word problems, without the complication of arithmetic algorithms, because the computer can serve as a calculator to handle those arithmetic operations that the tester does NOT want to emphasize. In an algebraic calculation, the computer can accept student inputs until some equation is achieved, then the computer can step in and solve the equation, then let the student make use of the solution in the following steps.

In short, the computer can "step in and help" at various steps in a procedure, just as a human teacher or examiner might: "Here, let me do that...Now, you can finish the problem."

19. The computer can attempt to correct misconceptions by using certain questions to identify the misconception, then re-wording things to clear the matter up. In this way the student can be led to respond to the intended problem, not to some other problem that the student mistakenly substitutes. [For example, if a student mistakenly interprets

$$\cos^{-1} 0$$

as

$$\frac{1}{\cos 0},$$

the computer can tell him that it really meant

$$\text{arc cos } 0.]$$

20. From the preceding possibilities, it should be clear that a computer can test for process and not merely for "products" or "results."

21. The computer can probe responses more deeply, as a human interrogator might do; in a sense, exploring a "tree" determined by student responses: "...and would that method always work?" "If k had not been positive, what would you have done?"

22. The computer can offer "hints," on any desired schedule, keeping track of which hints are needed, and perhaps "charging" for each hint.

The preceding were, in a sense, mainly from the point of view of the student. A computer can be very convenient for the teacher or examiner.

23. Since every student gets different numbers, and possibly even different questions, there should be less cheating.

24. Obviously, the computer will "correct" or "score" the test--no small thing!

25. The computer can report results in whatever form is most convenient.

26. Obviously, the computer is fast. Just like NBC on election nights the computer can report results just as soon as the vote is in, and it can report where desired, by locating terminals appropriately.

27. The computer can accept, and process, a very wide range of constructed responses. [If the answer is $V_2 - V_1$, the computer can easily recognize that

$$\frac{V_2^2 - V_1^2}{V_2 + V_1}$$

has the same value but is not in the simplest form.]

28. If desired, the computer can accept essay answers, and dispatch them to human readers, perhaps at remote locations, or retain them in memory for later scoring.

29. Improvability. In almost any way imaginable, the computer can collect data for the improvement of the tests (for example, by eliminating highly redundant questions) and can, if desired, even be programmed to collect data and make improvements automatically, without human intervention.

30. Teacher-made tests. The computer can contain a very long list of problems, from which a teacher can easily choose a selection appropriate to the teacher's class at any particular time, rather like ordering from a long menu in a restaurant.

31. For accountability or monitoring the computer can report immediately any indication of special problems, such as all the students in a class getting abnormally low scores (or abnormally high scores), or one question that nobody (or everybody) is answering correctly, or one school with scores that seem out of line, and so one.

BIBLIOGRAPHY

Anderson, Richard D., Jack W. Garon, and Joseph G. Gremillion. *School Mathematics Geometry*, Boston: Houghton Mifflin, 1973.

Bloom, Benjamin S. et. al. *Taxonomy of Educational Objectives. Handbook I: Cognitive Domain*, Longmans, Green, and Co., 1956.

Davis, Robert B., Elizabeth Jockusch, and Curtis McKnight. "Cognitive Proceses in Learning Algebra" in the *Journal of Children's Mathematical Behavior*, Vol. 2, No. 1 (1978).

Erlwanger, Stanley H. "Benny's Conception of Rules and Answers in IPI Mathematics," *The Journal of Children's Mathematical Behavior*, Vol. I, No. 2 (Autumn, 1973), pp. 7-26.

Erlwanger, Stanley H. *Case Studies of Children's Conceptions of Mathematics*. Doctoral Dissertation, University of Illinois, Urbana, Illinois, 1974.

Erlwanger, Stanley H. "Case Studies of Children's Concepts of Mathematics--Part I, *The Journal of Children's Mathematical Behavior*, Vol. I, No. 3 (Summer, 1975) pp. 157-283.

Hawkins, David. "Learning the Unteachable," Chapter I in the volume: Lee S. Shulman and Evan R. Keislar (ed.) *Learning by Discovery: A Critical Appraisal*. Rand McNally, 1966, pp. 3-12.

Hoffman, Banesh. *The Tyrany of Testing*, London: Crowell-Collier, 1962, New York: Collier Books, (1964).

Minsky, Marvin. "A Framework for Representing Knowledge," in: P.H. Winston, ed., *The Psychology of Computer Vision*. New York, N.Y.: McGraw-Hill, 1975.

Ortony, Andrew. "Remembering, Understanding, and Representation," *Cognitive Science*, vol. 2, no. 1 (Jan-March, 1978), pp. 53-69.

Rumelhart, David E.,, and Andrew Ortony. "The Representation of Knowledge in Memory," chap. 4 in: Richard C. Anderson, Rand Spiro, and William Montague, eds., *Schooling and the Acquisition of Knowledge*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977, pp. 99-135.

Stein, Sherman K. and Crabill, Calvin D. *Elementary Algebra: A Guided Inquiry*. Boston, Mass: Houghton, Mifflin, 1972.

Suppes, P. and Groen, G. (1967). "Some Counting Models for First-Grade Performance Data on Simple Addition Facts," in Joseph M. Scandura, ed., *Research in Mathematics Education*. Washington, D.C.: The National Council of Teachers of Mathematics.

Winograd, Terry. "A Framework for Understanding Discourse" in: Marcel Just and Patricia Carpenter, eds., *Cognitive Processes in Comprehension*. Hillsdale, N.J.: Lawrence Erlbaum Associates 1977, pp. 63-88.

INADEQUATELY TESTED ASPECTS
OF MATHEMATICS LEARNING

Leon Henkin
Department of Mathematics
University of California, Berkeley

with the cooperation of

Robert B. Davis
Curriculum Laboratory
University of Illinois, Urbana

Abstract

The information processing involved in mathematical work is decomposed into 8 categories, which may not all be distinct: knowledge of facts, mastery of algorithms, knowledge of concepts, ability to carry out meaningful meta-discussions where an algorithm itself becomes an object of discourse or analysis, knowledge of (and ability to use) heuristics, mastery of various "larger systems" (examples are given) and of a particularly important "larger system" that might be called the "language of mathematics," and, finally, "understanding". A joint discussion between the two authors has been presented in two papers, the present one (by L.H.) from the point of view of a logician, and the previous one (by R.B.D.) from the point of view of mathematics education.

If tests are designed to ascertain whether the aims of mathematics instruction are being realized, we ought to indicate what these aims are.

Of course as an empirical matter, the many mathematics instructors working in schools or other learning environments indubitably possess very diverse patterns of aims. The number of different sets of instructional aims would be still further increased by considering the motives that individual parents have for sending their children to school, or that individual public officials have for supporting mathematics instruction. However, in this paper we do not address ourselves to this vast set of aims. When, in this paper, we speak of "the aims of mathematics instruction," we have in mind simply aims that we ourselves propose, that we believe would have the support of many (but certainly not all) teachers and mathematicians, and for which we believe arguments can be adduced that will persuade substantial numbers of parents, public officials, and other concerned citizens, to give their support.

But even this specialized class of aims is only partially described here. One reason for the incompleteness of our description is our feeling that the project of uncovering aims is time-consuming and beset with subtle difficulties, and we have of necessity been confined to a limited effort in preparing this paper for an NIE-sponsored conference. Another limitation has to do with our variable opinions as to the degree and strength of support that we could muster for the various aims that we have been able to discern.

During several days early in July 1978 in Berkeley, the authors engaged in a thoroughgoing analysis of the components thought to merit exposition, and came up with the following list: algorithms, mathematical understanding, language and mathematics, knowledge of mathematical facts, concepts, and heuristics. As a convenient modus operandi, the first three were assigned to L.H. to write, and the latter three to R.D. Then, in a meeting in the East Asian Section of the Botanical Garden, an exploration was made of how these separately written portions could best be conjoined. Taking into account the purposes of this conference, it was felt that these would better be served if the individual emphases of the two authors were left intact.

The customary way to describe a mathematics curriculum is to give a list of "topics" constituting the "content", and to describe certain related "skills" which the student is expected to develop. A skill in this sense is a desired way of solving certain kinds of problems that arise in connection with the listed topics. Standard tests seek to ascertain (i) whether the student has knowledge of the content, by asking him to identify correct answers to questions about the topics of the curriculum, and (ii) whether the student possesses the desired skills, by examining his ability to solve problems of the kinds specified in the curriculum.

It is our position that the effect of mathematics instruction on students is much more complex than acquisition of the problem-solving skills and knowledge of the topics making up the curriculum. We contend that there are other aspects of mathematics learning that are inextricably interrelated with "course content" in the limited sense described above, which are crucial to successful performance in later courses (whether in mathematics or in related fields), which constitute a valuable part of the ultimate residue of mathematics learning that we hope will remain with an individual at the end of all formal studies, and whose assessment seems to be omitted from current, standard tests. We attempt to indicate some of these "other aspects" below.

Algorithms: Their place in mathematics and in school instruction

An algorithm is a uniform procedure, guaranteed to produce the correct answer to any problem of a specified class provided the problem is formulated in a standard way. Algorithms generally have a sequential character, and are described by means of rules indicating (i) how to start from the given problem in standard form, (ii) how to proceed at each intermediate stage according to specified features resulting from the preceding stage, and (iii) how to recognize when the procedure is finished and what the final answer is.

A mathematician tries to invent an algorithm when (s)he thinks (s)he recognizes a common pattern among solutions of several individual problems (s)he has solved. Among the mathematical activities undertaken in searching for an algorithm are the following: Seeking to define a large class of problems that will be easy to identify and amenable to solution by a common method; devising a notation for formulating these problems, which will lend itself to symbolic manipulation through a series of intermediate steps leading to a solution; estimating the "quantity of computation" involved in applying a procedure to a particular problem of the class; and seeking to replace complex or cumbersome (but effective) procedures by more efficient ones. Clearly activities of these kinds have potential application far beyond mathematics itself.

It is obvious that work with algorithms, but of a very different sort, makes up a large portion--probably it is fair to say the bulk--of the traditional elementary and secondary school mathematics curriculum. Starting with the algorithms for adding a pair of 2-digit numbers in first grade, and proceeding to high-school algorithms for solving equations, multiplying and factoring polynomials, or finding the measures of angles and sides of geometric figures, students are asked in each case to memorize the procedural rules of an alrorithm already found, and then to practice applying them to given problems so as to compute a substantial number of answers with few errors in a relatively brief period of time. It is these abilities that are the focus of current measurement by tests.

One of the important goals of the "new math" curriculum revisions that began in the mid 1950's was that of getting students to "understand" why the procedural rules of the traditional school algorithms give the correct solution to all problems in the class they are designed to solve. The traditional way in which a mathematician shows that (s)he understands the use of algorithms is by finding a proof that its use always leads to the correct solution of a given problem. Different proofs give different mathematical ways of understanding the same phenomenon. However, getting a student to repeat the steps of a proof that (s)he is "taught", is not convincing evidence that (s)he has the desired understanding; and expecting the student to find her/his own proof is generally far beyond her/his capability. Accordingly, new ways to communicate an understanding of why the traditional algorithms lead to correct solutions were sought--for example, through the use of specially designed manipulatives. But if understanding in this domain is taught and achieved through manipulating blocks, can it be adequately tested with paper-and-pencil answers to questions?

Questions of mathematical understanding are central to the aims of mathematics education. We defer to a later section of this paper a consideration of problems connected with defining the nature of understanding. But let us here indicate why it is desirable for students not merely to use standard algorithms to obtain numerical solutions to problems, but to understand why the numbers obtained by applying the algorithmic procedures constitute the desired solution. Of course for many persons understanding is an end in itself, and needs no other justification. But in the present context we want to argue that understanding is useful--to enable the student to develop tools for mathematical tasks (s)he may be called on to carry out at a later stage.

The fact is that students can well be asked to find certain algorithms, not merely to execute algorithms that are found by others. Of course the discovery of the basic arithmetic algorithms now in common use, of which the development of a place-value system of decimal notation was an essential part, was an enormous task whose working out spanned centuries. But within the framework of the major algorithms of arithmetic, algebra, and geometry that have already been devised and are now taught in schools, it is easy to identify particular classes of problems for which we may reasonably expect many students to devise their own algorithm.

For example, after solving a specific problem involving finding which point on a given circle is closest to a given line in a coordinate plane, students in a pre-calculus course can be asked to obtain an algorithm for finding the point of an arbitrarily given circle closest to an arbitrarily given line; they can go on to obtain an algorithm for finding the points within an arbitrarily given pair of ellipses that are closest to one another; and they can then be asked to find still more general classes of figures for which they are able to obtain algorithmic solutions to the problem of locating closest points in pairs of such figures.

The finding of algorithms involves the development of a host of valuable abilities such as were mentioned at the beginning of this section. But in the process of finding an algorithm, it is necessary to test hypothesized procedures to ascertain whether they do indeed lead to the desired solutions. And it is precisely to lay a groundwork for the ability to carry out such testing procedures, that it is desirable to ensure understanding, as well as competence in using, the standard algorithms that are taught at various stages in the schools.

Let us consider another example. Besides the standard algorithm for finding the product of a pair of 2-place numbers, there are of course many others that are widely used. For example, one can find the product of the multiples of 10 that are closest to the given numbers, and then make certain adjustments to this intermediate number. Such an alternative algorithm is actually quicker than the standard one for a certain limited class of pairs of numbers--but what is that class? How do we measure the "quickness" of an algorithm to find the product for a particular number pair? These are very reasonable questions to discuss with students, and to set them investigating. Obviously, however, the student who has not passed beyond efficient use of the standard algorithms, to an understanding of them, will be greatly handicapped in such investigations.

We have not yet mentioned the appearance of calculators and computers in schools and homes. Increasingly, these will take over the function of executing algorithms of all sorts, and the demand on humans will be to devise the algorithms whose execution is then delegated to the machine. It is impossible to devise an algorithm unless one can determine whether a given procedure does or does not lead to correct solutions for problems of a given class. We must begin by understanding why the standard algorithms succeed.

There are parts of mathematics where we have several algorithms that are applicable to classes of problems that overlap, and where the question as to which of these classes contains a given problem may not be very evident. A good example is the integration of elementary functions, where a handful of standard techniques (e.g., various kinds of substitutions, integration by parts) are each applicable to some functions, while other elementary functions are known not to possess an elementary antiderivative at all. To be successful in using such a set of algorithms requires skills and understanding that pass beyond those acquired by memorizing procedures and practising their use.

Of course memorization of procedures will always be a bedrock component for a frequently used algorithm. But as we come to understand more about the way in which related items are stored in memory, there is increasing evidence that memorization of procedures of any substantial complexity is itself greatly aided by understanding how and why the components of the algorithm fit together.

Problems susceptible of solution by algorithmic computation are generally contrasted with "word problems" that are found in all school mathematics books and courses; it is the latter on which many students founder. A large part of the difficulty arises from the fact that the word problems are not presented in that kind of standard symbolic form upon which our customary algorithms are fashioned to operate.

Actually, it is easy enough to devise an algorithm for many classes of word problems to be found in text books. Consider, for instance, the following sequence of sentence forms.

> "J is p times as old as M. In y years, J will be
> q times as old as M. How old was M z years ago?"

Let C be the class of problems obtained by sustituting the first names of people for the variables "J" and "M", and substituting numerals for the variables "p", "q", "y", and "z" in this sequence. Problems in this class C can be found in many current textbooks; to devise an algorithm for solving an arbitrarily given problem in the class C is certainly within the capability of some students.

It is common practice among mathematics educators and cognitive psychologists to use the word "strategy" in connection with a procedure for seeking the solution of a problem of a certain kind. In some cases a strategy, in this sense, is nothing other than an algorithm; but for the most part it differs in one or another way. Often, the problems to which a given strategy is applicable are not formulated in a completely standard, symbolic way; often, when a particular intermediate step is reached through following the strategy, the procedure leading to the next following step is not completely prescribed, but instead several possible procedures are listed and it is left to the intuitive judgment of the user to select which procedure is used next; and sometimes--most disappointing!--the sequence of intermediate steps does not terminate with the desired solution, but just "peters out", leaving the user to make a fresh start. In some cases it would be possible to replace such a strategy with a true algorithm, but it is not clear that this is desirable from the point of view of promoting learning--even though it might result in a higher ratio of problems correctly solved. Students often fail to distinguish between a strategy and a true algorithm, and are disappointed when a strategy fails to perform like an algorithm, not realizing that this failure is not their own personal failure.

What is the role of standardized testing with regard to algorithms? Obviously it is desirable to test more than the memorization of procedural rules and the swiftness and accuracy of their application. One desideratum would be to test the student's ability to devise a new algorithm for a simple class of problems--but the limited time generally thought to be available for standard tests makes this difficult to implement. At the very least, however, we should set ourselves the goal of testing the extent to which a student understands why the algorithms (s)he uses lead to correct solutions. But how do we do this?

Let us examine the process of mathematical understanding in a broader context.

Mathematical understanding

Although the generation of understanding is widely acknowledged to be a central aim of education, a satisfactory definition of mathematical understanding is not available. It is evident that one difficulty in obtaining such a definition has to do with the multitude of components of mathematical understanding, and the rich way in which these are intertwined. The intellectual abilities identified in Bloom's "Taxonomy of Educational Objectives"--comprehension (the "lowest level of understanding", apprehension of what is being communicated), application, analysis, synthesis, and evaluation--may all be involved, yet there are certainly elements of mathematical understanding that seem to lie beyond all of these.

Let us consider a simple example. At the age when young children are beginning to learn counting, a widespread phenomenon is boasting about the high level to which one is able to count. At a certain stage a child is proud that (s)he is able to count to 10; later to 30 or 40; still later to 100. And then, as the learning process proceeds, this pattern of prideful expression stops: the child comes to understand that a claim to be able to count to a given level is not so much an assertion of achievement, as an indication of non-achievement (of ability to count beyond that level). That is, (s)he has grasped an essential part of the concept of the infinitude of the counting numbers.

How can we test whether such comprehension is possessed by an individual student at a given time? Of course if her/his ability to count is in fact limited by a certain number n, one can discover this by asking her/him to count beyond n--providing the fact (s)he fails in this task can be reliably judged not to result from a unwillingness to accommodate the questioner! But if in fact the child does have the understanding being tested, in principle no specific counting task can reveal this.

On the other hand, one certainly cannot obtain the desired information about the child's understanding of this point simply by asking, "Is there a highest number to which it is possible to count?"--because it is easy for a teacher to tell a student that the answer to this question is no, and to drill the student to memorize the answer and give it in response to the question on a test. Yet knowledge (or belief) that the answer is no, acquired by being told, is not in itself the understanding about which we are writing. The essential points are the realization by the child that no matter what large numeral (of many places) he may be presented with, he has the capability of producing the next larger numeral--together with the knowledge that this production of numerals (rather than the use of special words like "billion" and "quadrillion") is the essence of counting ability. On the other hand, the description of the general rule for obtaining the next numeral from an arbitrarily given one is sufficiently complex so that very few students have the capability of formulating it until long after they have the particular mathematical understanding about which we are writing.

The normal way to test mathematical aptitudes as well as achievements is through problem-solving. Cognitive psychologists have turned their attention increasingly to an analysis of problem-solving performance; a promising mode of this analysis is in terms of computer-program modeling, using concepts from theories of artificial intelligence. However, although problem-solving is certainly abetted by an understanding of the concepts and methods involved in the problems and their solutions, it is certainly possible to solve large classes of problems without understanding. We would say that computers exemplify this contention--though we realize this is a controversial subject!--and, unfortunately, its exemplification by students is only too familiar.

Note that the particular understanding about the nature of counting, which we have contended above arises naturally in many students at a certain stage in their work with counting, is not of use in helping them to solve any of the usual counting problems with which they are presented--and hence, as we have noted, the understanding cannot be tested by posing such problems. This is not to say that the understanding is useless--on the contrary, an understanding of the infinitude of the counting numbers is absolutely essential to deal adequately with the mathematical notions of limit and continuity that underlie the differential and integral calculus. But of course problems connected with these concepts can hardly be used to test for the understanding at the time it arises in young children.

Perhaps the example about counting that we have been considering may be considered too special, an isolated curiosity. Some may even consider the learning of counting to be a pre-school activity and so "hors-de-combat" as regards discussion of standardized testing. But the fact is that the observations that we have formulated above can be made and reinforced at almost every point of the mathematics curriculum. Recently one of us (L.H., with students David Ferguson, Hadas Rin, Philip Treisman, and Alan Sanstad) has been investigating various areas where mathematical understanding, or lack of it, impinges in a serious way on school mathematics learning. One area to which our attention has been drawn is the subject of quadratic equations, which rightfully occupies a solid position in the secondary mathematics curriculum. The ability of many--but not all--students to achieve substantial success in solving a good proportion of the problems normally presented in this area is familiar. Yet our observations suggest that even the very good students often have a relatively meager understanding of the concepts involved. In Appendix Q below we shall indicate some of the elements that we consider are involved in this kind of understanding.

While we are far from being able to give a methodologically-acceptable definition of mathematical understanding, even for a specific area such as counting or quadratic equations, we think we can identify some important elements, and we can point to some important obstacles to testing for the essential components of understanding.

1. Mathematical understanding seems to involve an organization of some area of mathematical experience, which results in a capability of connecting it in a mathematically significant way with quite new areas of mathematical experience not yet encountered.

2. One can never test understanding of concept A by subject S, through observing whether S can solve problems relating A to a different mathematical area B, unless one knows whether B is "new" to S. That is, the same response that an S_1, who truly understands A, will make to a problem involving B, can be achieved by an S_2, who lacks this understanding, through advance "training".

3. Mathematical understanding in a given area is connected with an ability to recognize and distinguish between steps that are "legal" and steps that are "desirable" in achieving a given aim--computation, construction, or proof.

4. Mathematical understanding of a given phenomenon involves not merely knowing the facts involved, but also possessing insight into "why" the facts are as they are. This usually means a perception of how the facts in question are related to a larger, previously known body of facts. As there may be many different relations of this kind, different "ways of

understanding" the same phenomenon are possible; also, individual S_1 will have a "greater understanding" of phenomenon P than S_2, if S_1 perceives all of the relations that are known to S_2 and others besides.

5. To test for understanding of the kind just described, it is natural to ask for an *explanation*. However, if a student furnishes some words by way of explanation, there is a possibility that they do not express an insight possessed by the student, but simply constitute an explanation that was furnished by a book or a teacher without the explanation being understood by the student. On the other hand, if the student fails to furnish an explanation, it is still possible that (s)he possesses the insight constituting understanding, but lacks the verbal facility to formulate it for communication.

This brings us to a consideration of the linguistic aspect of mathematics learning.

Added note After the above section of the paper was completed, our attention was called to the booklet "Objectives for Mathematics Learning" by Shmuel Avital (assisted by Sara Shettleworth), published in 1968 by the Ontario Institute for Studies in Education. In this work Avital takes the categories of educational objectives listed in Bloom's *Taxonomy*, and attempts to modify and then illustrate the classification in the field of mathematics education. Avital distinguishes three "levels of thinking", relating them to Bloom's categories of objectives as follows. (1) "Recognition and recall" is employed to demonstrate the category of Knowledge; (2) "Algorithmic thinking and generalization" are employed in meeting objectives falling under Comprehension and Application; (3) "Open search", the highest level of thinking, combines Analysis and Synthesis. The sixth of Bloom's categories, evaluation, is found by Avital to be "psychologically indistinguishable" from Analysis and Synthesis in the context of mathematical performance. Avital distinguishes his third level of mathematical thinking from the second by the number of algorithms employed, by the "remoteness" of these algorithms from the clues that are given in the problem, and by the remoteness from one another of the algorithms that are combined. Avital's analysis of mathematical thinking as having components on three levels does not deal *explicitly* with the mechanism which leads to the finding of algorithms remote from given clues, or from each other, on the third level. Presumably this is generated by a sort of insight, which for us is the essence of understanding. Indeed, it is a familiar experience for a mathematician to find a solution to some problem and yet to feel that (s)he doesn't understand the solution. (This feeling is sometimes, but not always, engendered by observing that a problem thought to be "essentially similar" to the first cannot be solved by the procedures that led to the original solution.)

Language and mathematics

Comprehension, in Bloom's *Taxonomy*, is the lowest level of understanding--it is simply the student interpreting the words being used by the teacher or textbook or test so as to apprehend their meaning. We have to realize, however, that this form of understanding is beset by special difficulties in the realm of mathematics.

A teacher asks, "What do the following sets have in common: the set of integers less than 5, and the set of integers greater than 1?" (S)he expects the answer "2, 3 and 4," but the student replies "Both are sets of integers." In the context of the lesson on sets and numbers, the teacher interprets the word as referring to properties of sets. The student has misunderstood the teacher's question, but has shown a good understanding of the relation between the sets in question. Whether the teacher recognizes this depends on whether (s)he is sensitive to the possible reasonable interpretations of the question--aside from the interpretation that was intended.

Miscommunications of this kind are especially prevalent in the context of mathematics learning, for several reasons. For one thing, not only do special terms abound, but also ordinary words are sometimes given a technical meaning somewhat at variance with the familiar one. For example, in modern algebra the term "greatest common divisor" refers not to that common divisor which is greater than all others, but rather to any common divisor (there are often more than one!) which is a multiple of all others. Perhaps even more confusing is the fact that ordinary words which do not appear to be technical terms at all are used within mathematical discourse in ways that are slightly out of the ordinary. For example, the word "or" is normally used in a way that permits changes between the exclusive sense ("one or the other but not both") and the inclusive sense ("one or the other or possibly both") from one context to another, whereas in mathematics "or" is used only in its inclusive sense.

An especially confusing terminological phenomenon is the double use of the word pair "if ... then", in mathematics, to express an implication between two sentences and to form a conditional sentence for a given pair of sentences. For example, the conditional sentence "If 3 is a factor of 5 then 3 is a factor of 10" would be avoided in ordinary discourse if one knows that 3 is not a factor of 5. Yet it is a sentence that is not only mathematically acceptable, but one that must be considered true. This is because of its relation to the patently true sentence "Every factor of 5 is a factor of 10", which can equivalently be expressed in the mathematical form "For every integer x, if x is a factor of 5, then x is a factor of 10". The original sentence referring to 3 is seen to be simply a special case--an instantiation--of this universal truth. In turn, this universal sentence is a consequence of the law "For all integers x, a, and b, if x is a factor of a, then x is a factor of a x b" together with the fact that 10 = 5 x 2. And the law, in turn, is established by showing (with the aid of the associative law of multiplication) that "x is a factor of a" implies "x is a factor of a x b".

The use of variables has crept into the preceding discussion of the "unnatural" use of the words "if ... then" in mathematics. In fact, misunderstanding of the use of variables in mathematical language is by far the most widespread source of confusion among students. A nice example is given in Avital's "Objectives for Mathematics Learning", p. 55.

> Jan says that (A) an integer divisible by n can be written as nt where t is an integer; (B) $3t$ represents integers divisible by 3 and $7t$ represents integers divisible by 7; (C) $3t + 7t = 10t$; hence (D) the sum of an integer divisible by 3 and an integer divisible by 7 is always divisible by 10.

Avital poses the problem whether Jan's argument is right or wrong, and in the latter case the student is asked to indicte whether this mistake lies in (A), or (B), or the passage from (B) to (C), or the passage from (C) to (D).

What makes this problem so difficult to undertand for most students is the mathematico-colloquial confusion of use and mention of symbols in (A) and (B), and the hidden use of variables in (D) that is signaled by the twice-used indefinite article "an" and the word "always". Let us explain.

The basic function of letters in mathematical language is to refer to the objects of mathematical discourse--in our example, integers. In principle, there is no need to *refer to the letters* at all; we *use* the letters to *refer to numbers*. But when, in (A), we say that a certain integer "can be written as nt", we seem to be *mentioning* the letters, indicating that we write them to refer to the integer. Similarly, we are talking *about* the expression "$3t$" in (B) when we say that it "represents" a certain kind of integer. The proper formulations of (A) and (B) are as follows. (A') An integer a is divisible by n if and only if $a = nt$ for some integer t. (B') If a is divisible by 3, then $a = 3t$ for some integer t, and if a is divisible by 7 then $a = 7t$ for some integer t.

That phrase "for some integer t" that apears in B' and is left out of (B) "as a convenient shorthand", is actually the key to understanding why Jan made a mistake in his argument. Let us see why.

To express (D) by means of variables we use (D'): "For any integers, a and b, if a is divisible by 3 and b is divisible by 7, then $a + t$ is divisible by 10." Now if Jan were seeking to prove (D') he would start by writing, "Choose any integer a divisible by 3 and any integer b divisible by 7. Let us try to show that $a + b$ is divisible by 10. Well, by (B') we know that $a = 3t$ for some integer t--say $a = 3u$; and $b = 7t$ for some integer t--say $b = 7v$. Thus $a + b = 3u + 7v$." And now, of course, it would be clear to Jan that (C) is of no use in seeking to conclude that $a + b = 10t$ for some integer t.

The teacher understands that in trying to use (B) to prove (D) (s)he must use an expression with 2 variables such "$3u + 7v$", and not an expression with a single variable such as appears in (C); but the beginning student often does not. We suggest that the understanding of the teacher in this case is not at all a deeper grasp of the concept of divisibility of integers, but is simply the linguistic facility to interpret (B) as meaning what (B') says, and the skill to employ the rule that the phrase "for some t" requires the introduction of a new letter at each application.

When we speak here of "linguistic facility" do we mean grammar? Or are we really speaking about logic--a portion of mathematics? Most people are unaware that there are now mathematical theories of grammar as well as of logic, and that the fundamental concept of mathematical logic is that of a formal deductive system--which is based on an idealized language having its own recursive grammatical rules. But in fact we are not speaking of either grammar or logic when we refer to the difficulties that students have in comprehending the langugage of mathematics--and certainly we are not advocating the study of these disciplines as a way of overcoming these difficulties. Essentially, the language used in (B) and (D) to express the respective meanings of (B') and (D'), in the above example, is a kind of "mathematician's slang." You do not become adept at using slang by studying grammar or logic--you have to be around people who use the slang regularly, and then you have to get practice in using it yourself.

The need to introduce new variables in each application of a phrase such as "for some integer t", which was illustrated in our analysis of the example above, is a linguistic skill whose importance is cumulative as a student progresses in the learning of mathematics. Yet it is never listed as one of the "topics" forming the subject matter of a mathematics course, and so it remains untested: If the student messes up an argument in the way that Jan did in Avital's example, (s)he is judged to be deficient in solving problems on divisibility instead of lacking in linguistic skills.

The tendency of mathematical slang to shorten discourse by dropping prases such as "for some integer t", or "for all numbers x", is an important source of misunderstanding in elementary algebra. Equations, instead of serving as parts of sentences to communicate about numbers, are taken out of the sentential context and are themselves made objects of study and symbolic manipulation. Thus, when the teacher says "Solve the equation $x + 7 = 10$" (s)he appears to be talking about the symbolic expression "$x + 7 = 10$", but is "really" using this expression to talk about a certain number--namely, that number x which, when added to 7, results in the number 10.

And when, in response to the teacher's request, the student writes the equation "$x + 7 = 10$" followed by "$x = 10 - 7$"--neither equation viewed as a part of a sentence and with no words to indicate the relation between the two--(s)he is often unaware that this is shorthand for expressing the idea conveyed by the sentence, "If x is any number such that $x + 7 = 10$, then also $x = 10 - 7$." For such simple equations as here considered this hardly seems to lead to confusion, but already in the case of quadratic equations, and especially in the solution of equations involving surds, a failure to appreciate these linguistic transformations is the root of deep confusion for many students. We have expatiated on these ideas in our paper "Linguistic Aspects of Mathematical Education," which appears in "Learning and the Nature of Mathematics" (SRA, 1972) edited by William E. Lamon.

Of course, in the end, we want our students to be able to manipulate equations in the search for numerical solutions as well as in the construction of proofs. But they will not understand what they are doing unless they realize that such manipulation is a linguistic shorthand for sentences involving phrases such as "for some integer t", "for all numbers x", and "if ... then", which do not actually appear in the manipulations. For the teacher, these observations suggest the task of using the full sentences--and training the student to use them--when the ideas are first discussed, and then gradually moving to the shorthand format in which equations stand alone. For the testmaker, the observations we have made suggest the task of testing whether indeed the students possess the skills of passing back and forth between equivalent linguistic forms for expressing mathematical ideas.

In the preceding paragraph we have spoken of the lack of certain linguistic skills as impeding mathematical understanding. It is probably widely recognized that in general language skill is closely related with understanding; but in the context of mathematics the two notions are so closely bound to one another that it may be defensible to identify them. Because of the mathematician's propensity to make symbolic manipulations the object of study as well as the means, mathematics itself becomes language. This is recognized in the familiar dictum that "mathematics is the language of science," as well as in the logician's treatment of a mathematical theory as an idealized language described by transformational rules. Thus, to test mathematical understanding may come to be regarded as "nothing but" testing the ability to use mathematical language.

Appendix Q

The topic of quadratic equations (q.e.'s) is solidly ensconced in the secondary mathematics curriculum. In this area the course focuses on finding solutions to given q.e.'s by means of one of several algorithms. However, a good understanding of this subject involves awareness of a great many aspects of, and relations among, the steps in the algorithms, as well as perception of relations linking q.e.'s with graphs, with functions, with other classes of mathematical statements, and with non-mathematical phenomena to which q.e.'s can be applied. Here is a partial list of skills, knowledge, and understandings that are connected with this subject.

1. Ability to furnish examples of q.e.'s, and then to identify which ones, in a given list of equations with numerical coefficients, are quadratic; ability to describe the totality of q.e.'s by means of parameter coefficients.

2. Ability to use an algorithm for converting q.e.'s to "standard form" (first with numerical coeff's, then with coeff's that contain parameters).

3. Understanding what it means for a number to be a root of a q.e. This includes knowledge that: (a) equations with the same coefficients but different variables have the same roots; (b) a given q.e. and its standard-form equivalent have the same roots; (c) a q.e. with parameter coefficients does not, in general, have numerical roots, but when numerical values are assigned (in any way) to the parameters, this determines numerical roots. [Item (c) is part of a general understanding that the purpose of using parameters in any mathematical statement is to consider a variety of special values.]

4. Understanding the relation between the roots of a standard-form q.e., and the places where the graph of the left side of the equation crosses the horizontal coordinate axis. Perception of the graphical significance of replacing the right side of a standard-form q.e. by a constant other than zero, and ability to predict how such a change affects the roots.

5. Knowledge that the number of (real) roots of a q.e. may be either 0, 1, or 2--and that there are no other possibilities. Understanding this phenomenon in two different ways, as follows: (a) by relation to crossing of the horizontal axis by the graph of the q.e. (presupposing knowledge and understanding about the form of such graphs); (b) by relating the roots of a standard-form q.e. to the factorization of its left side. Note that (b) involves background understanding about factoring formal polynomials and its relation to factoring numbers, and depends on the no-zero-divisor property of the real numbers. The employment of this principle requires an inference from the sentence "If $(x-a)(x-b) = 0$ then either $x - a = 0$ or $x - b = 0$" to the sentence "There are no roots of $(x - a)(x - b) = 0$ other than a and b"; but the hypothesis

does not imply that either a or b is a root of the given equation. There must be a separate understanding as to why a, b are roots of above q.e., and then the two facts are conjoined to conclude that a, b is a complete list of roots. There should be understanding of the use of the words or and and in the following pair of equivalent sentences: (i) x satisfies $(x - a)(x - b) = 0$ if and only if $x = a$ or $x = b$, and (ii) the roots of $(x - a)(x - b) = 0$ are a and b.

6. Ability to apply each of the three standard algorithms to find the solution(s) of appropriate q.e.'s: factoring, completing the square, using the quadratic formula. Knowing when factorization with integer coefficients is possible; knowing the limitation in using the quadratic formula to obtain real roots; understanding the graphical significance of the equation obtained by completing the square; possessing a good intuition as to the order in which to try the algorithms in specific cases; understanding that the quadratic formula results from using completion-of-square procedure on a q.e. with parameter coefficients, and that the quadratic formula leads to a formal factorization of such a q.e.

7. Ability to obtain a standard-form q.e. possessing given roots; knowledge of relation between coefficients of q.e. and sum, product of roots; ability to give parametric description of q.e.'s possessing only one root, or with two roots in a given relation (i.e., one the negative of the other, or twice the other).

8. Ability to indicate qualitatively the effect on the roots of a given standard-form q.e. of a given change in the constant term.

9. Understanding the relation between the graph of a linear function and the graphs of quadratic functions--the former a limit of the latter.

10. Ability to set up q.e.'s in order to solve a variety of standard word problems; understanding why area problems often lead to q.e.'s, and why some max/min problems lead to q.e.'s but never to linear equations; understanding the relation between the max or min of a quadratic function, and the roots of the corresponding q.e.

Many of the skills and classes of facts listed above can be used in testing understanding--but only if the tester has prior knowledge of the degree to which the student is familiar with the concepts or vocabulary involved. To take an extreme example, we encountered a student who could not respond properly to the request "Can you write down an example of a quadratic equation", yet who was able to use without hesitation two of the standard algorithms to obtain the correct roots of q.e.'s that were given to her, and was familiar with the term "quadratic formula." It is a fair inference that the term "quadratic equation" was mentioned in this student's class when the topic was first considered, but it was not a term actively employed in subsequent class work, and hence did not remain a part of her active vocabulary.

MATHEMATICS TESTING: A VIEW FROM THE SCHOOLS

Ross Taylor
Mathematics Consultant K-12
Minneapolis Public Schools

Introduction

The principal purpose of testing is to provide information to the educational decision makers at all levels. For example, students and teachers need information for diagnosis and for monitoring achievement. Guidance counselors need information to help students plan for their futures. Administrators need information for administrative decisions.

Ideally, information gained from testing will be used for the improvement of teaching and learning. Realistically, political considerations and financial constraints are the major factors that influence administrative decisions. In today's environment these two factors are having increasing significance.

Today, school costs are escalating due to inflation and to the maturing of faculties as a result of declining enrollment. Enrollment decline also means a decrease in the per-pupil allocated dollars from the states. Dollars from local taxes are getting even more scarce. The passage of Proposition 13 in California is a signal that taxpayers are fed up with taxes in general and property taxes in particular. The effects of this so-called "taxpayer revolt" are being severely felt by school systems throughout the country.

Today, the public is making increasing demands for accountability by the schools. The rallying cry is "back to the basics" and the result is a headlong rush into minimum competency testing. In the May, 1978, issue of Phi Delta Kappan, Chris Pipho of the research department of the Educational Commission of the States reported:

> What began as a startling idea in California, Florida, Oregon, and a handful of other states in 1975 and 1976 has now arrived in some form in each state. As of March 15, 1978, 33 states had taken some type of action to mandate the setting of minimum competency standards for elementary and secondary students. All the remaining states either have legislation pending or legislative or state board studies under way.
> (Pipho, 1978, p. 585)

Initial minimum competency programs were often hastily conceived and implemented with the naive assumptions that higher achievement could be legislated, that no special funds were needed for testing and remediation programs, and that suitable tests were readily available. The minimum competency movement is maturing as experience is gained and as the activity moves from legislatures to state departments of education to the schools. Action at all three levels is becoming more deliberate and considerate as the complexities of testing and the other issues emerge. The time is ripe for investigation, development and implementation of appropriate testing models.

As schools face the external forces of limited funding and demands for accountability, they have an urgent need to focus on a systematic development of cost-effective programs of instruction in the basic skills. A systematic approach must begin with a determination of which skills are basic and the identification or development of evaluation programs to measure achievement of these skills.

What Are the Basic Mathematical Skills?

The National Institute of Education (NIE) sensed a need to determine which mathematical skills are basic, so it sponsored a three-day Conference on Basic Mathematical Skills and Learning held in Euclid, Ohio, in October, 1975. Invited participants contributed 33 papers addressing the following questions:

1. What *are* basic mathematical skills and learning?

2. What are the major problems related to children's acquisition of basic mathematical skills and learning, and what role should the National Institute of Education play in addressing these problems?

(NIE, Vol. I, 1975, p. vii)

At the Conference four working groups were formed to make recommendations in the following areas:

1. Classroom Instruction and Teacher Education
2. Curriculum Development and Implementation
3. Goals for Basic Mathematical Skills and Learning
4. Research Priorities

The report of the working group on Goals for Basic Mathematical Skills and Learning states the following general goals:

> The purpose of mathematics education is threefold: (1) to prepare the child for life as a consumer, voter, and citizen; (2) to begin the training for a variety of productive occupations and professions; (3) to assist the child in developing a rich and rewarding life.
>
> (NIE, Vol. II, 1975, p. 17)

In May, 1974, the Conference Board of the Mathematical Sciences appointed a National Committee on Mathematical Education (NACOME) to prepare an overview and analysis of U.S. school-level mathematical education--its objectives, current practices, and attainments. The NACOME report, published in the fall of 1975, states:

> The questions to be answered by investigation of achievement and attitude data are deceptively straightforward:
>
> (1) Are students acquiring the mathematical skills and understanding needed for meaningful participation in our contemporary technological society?
>
> (2) Is school mathematics successful in preparing interested students for careers in business, industry, government, engineering, and the sciences like physics, biology, statistics, or computers?
>
> (3) Are students developing positive attitudes toward the methods of mathematics and appreciation of the role that mathematics has played in development of contemporary culture?
>
> (NACOME, 1975, p. 102)

The similarity of the statement of purpose in the Euclid Conference Report and the questions to be answered by investigation in the NACOME report is striking. There appears to be validity in the consideration of mathematics in the schools with respect to: (1) basic skills for all citizens; (2) mathematics for careers; and (3) affective and cultural aspects of mathematics. Today, the pressure on the schools is primarily in the area of basic skills for all citizens. The Euclid Conference working group listed and described ten basic goals in mathematics for all citizens and five further desirable goals. "Appropriate Computational Skills" is one of the ten basic goals in the Euclid Conference report. There was a universal concern at the conference, expressed in many of the contributed papers, that the public, as well as some educators, tend to focus primarily on computational skills and neglect the other basic mathematical skills.

In April, 1976, a special session of the annual meeting of the National Council of Supervisors of Mathematics (NCSM) was held to discuss the Euclid Conference Report. Supervisors expressed the concern that they were under pressure to concentrate on computational skills and they were afraid that other basic mathematical skills would be neglected. They called for NCSM to develop a unified position on basic mathematical skills that would give them needed ammunition in their battle against a narrow concentration on computation. With funding from NIE, an NCSM task force met in Minneapolis in June of 1976 to prepare a first draft of a Position Paper on Basic Mathematical Skills. Reactions to the first draft were solicited from the membership of NCSM and the suggestions received were incorporated into the final paper which was published in January, 1977. The NCSM Position Paper lists the following ten basic skills areas:

Problem Solving
Applying Mathematics to Everyday Situations
Alertness to the Reasonableness of Results
Estimation and Approximation
Appropriate Computational Skills
Geometry
Measurement
Reading, Interpreting, and Constructing Tables, Charts, and Graphs
Using Mathematics to Predict
Computer Literacy

(NCSM, 1977)

The NCSM task force used the ten Euclid Conference basic goals as a starting point, so it is not a coincidence that the two lists are very similar. The NCSM paper added Geometry as a topic and did not include the Euclid Conference topic of Quantitative Understanding of and Drawing Inferences from Functions and Rates of Change in its final list. The task force made a number of changes and modifications in the descriptions of the other nine Euclid Conference Basic Goals. The Euclid Conference headings of Problem Solving, Alertness to Reasonableness of Results, Estimation and Approximation and Appropriate Computational Skills were included in the NCSM list. The other headings were modified. For example, Links Between Mathematical Ideas and Physical Situations was changed to Applying Mathematics to Everyday Situations and Notions of Probability was replaced with Using Mathematics to Predict.

The NCSM Position Paper has received almost universal acceptance by mathematics educators. It has been endorsed by the Delegate Assembly of the National Council of Teachers of Mathematics (NCTM) and by the NCTM Executive Board. It has been widely disseminated, including publication in the NCTM journals and a number of other educational publications. Textbook publishers, test publishers and vendors of other educational materials have used the NCSM Position Paper as a reference for producing educational Materials. For example, the Educational Testing Service used the NCSM basic skills list as a guideline for developing its basic mathematical skills testing program.

While the NCSM list offers a general framework for basic skills, it does not list specific objectives, nor does it contain sample test items to measure attainment of these objectives. There is much to be done in the development of specific basic skill objectives and tests to measure achievement of these objectives. In areas such as computation and measurement, clearly defined objectives, test items and instructional materials are available. However, considerable development is needed in areas such as Using Mathematics to Predict and Computer Literacy.

There should be a clear distinction between the terms basic skills and minimum competence. Basic skills should be interpreted as broadly as possible. In particular, in mathematics it must not be restricted to computation. Minimum competence should be used as a relative term with specific meaning only in a specific context, such as the minimum competence necessary for graduation from a particular school. Universal agreement on minimum competence necessary for all citizens is not likely to be achieved. If minimum competence is to be defined, it should be done at the local level. Then if the distinction between basic skills and minimum competence is maintained, the broad concept of basic skills will not be narrrowed or compromised.

For an example of how the NCSM basic skills can influence the curriculum, consider Minneapolis. The Minneapolis Schools have chosen not to go the minimum competence route. Instead, a basic skills test keyed to the ten NCSM basic skills areas is given to tenth graders. The results are used to counsel students who do not do well to take more mathematics. Curriculum development and staff development are focusing on areas of weakness as indicated by item analysis.

Basic Mathematical Skills and Testing

The NCSM Position Paper on Basic Mathematical Skills contains the following section on Evaluating and Reporting Student Progress:

> Any systematic attempt to develop basic skills must necessarily be concerned with evaluating and reporting pupil progress.
>
> In evaluation, test results are used to judge the effectiveness of the instructional process and to make needed adjustments in the curriculum and instruction for the individual student. In general, both educators and the public have accepted and emphasized an overuse of and overconfidence in the results of standardized tests. Standardized tests yield comparisons between students and can provide a rank ordering of individuals, schools, or districts. However, standardized tests have several limitations including the following:
>
> a. Items are not necessarily generated to measure a specific objective or instructional aim.
>
> b. The tests measure only a sample of the content that makes up a program; certain outcomes are not measured at all.
>
> Because they do not supply sufficient information about how much mathematics a student knows, standardized tests are not the best instruments available for reporting individual pupil growth. Other alternatives such as criterion tests or competency tests must be considered. In criterion tests, items are generated which measure the specific objectives of the program and which establish the student's level of mastery of these objectives. Competency tests are designed to determine if the individual has mastered the skills necessary for a certain purpose such

> as entry into the job market. There is also need for open-ended assessments such as observations, interviews, and manipulative tasks to assess skills which paper and pencil tests do not measure adequately.
>
> Reports of pupil progress will surely be made. But, while standardized tests will probably continue to dominate the testing scene for several years, there is an urgent need to begin reporting pupil progress in other terms, such as criterion tests and competency measures. This will also demand an immediate and extensive program of inservice education to instruct the general public on the meaning and interpretation of such data and to enable teachers to use testing as a vital part of the instructional process.
>
> Large scale testing, whether involving all students or a random sample, can result in interpretations which have great influence on curriculum revisions and development. Test results can indicate, for example, that a particular mathematical topic is being taught at the wrong time in the student's development and that it might better be introduced later or earlier in the curriculum. Or, the results might indicate that students are confused about some topic as a result of inappropriate teaching procedures. In any case, test results should be carefully examined by educators with special skills in the area of curriculum development.
>
> (NCSM, 1977)

The inappropriateness of standardized norm-referenced tests for individual student assessment and program evaluation is discussed in the NACOME report.

> Norm-referenced tests are widely used both for individual student assessment and for program evaluation. As an instrument for individual student evaluation, the test provides a score (or possibly two or three subscores) which compare his performance with that of the norming population. Furthermore, the score (or scores) represent average performance in a wide range of content areas at several cognitive levels. Hence, areas of relative strength and weakness cannot be determined. This fact makes the test practically useless for diagnostic purposes and thus a poor instrument for evaluating student progress toward individual educational goals.
>
> The standard norm-referenced tests are even less appropriate for program evaluation. Only rarely are programs directed solely at an overall improvement in student performance. Usually the program has a specific set of objectives and testing should focus on these objectives.
>
> (NACOME, 1975, pp. 123-4)

The problem of finding suitable tests for evaluating compensatory education programs is noted in a report of the compensatory education study staff of NIE.

> Programs are typically evaluated using national norm-referenced tests which are not designed to identify and discriminate among different

> skills within a subject area. However, criterion-referenced tests which are more appropriate for comparing different educational procedures, are not yet well developed technically.
>
> (Rotberg and Wolf, 1974, p. 41)

Standardized norm-referenced tests are particularly inappropriate for minimum competency programs. In order to discriminate among students, they must contain many items that half or more of the students will miss. However, in a minimum competency program the tests should consist of items that most students should be able to answer correctly after receiving suitable instruction.

Standardized norm-referenced tests are clearly inappropriate in the basic skills areas and criterion-referenced tests appear to show more promise. However, as criterion-referenced tests gain more widespread use, problems such as a tendency to focus on precisely defined isolated skills may emerge. Various alternatives should be investigated. For example, the NACOME report suggests the development of short tests which are directed at specific content areas and normed on suitable populations.

(NACOME, 1975, p. 125)

The call for research and development in testing is heard again and again. The Report of the Conference on Needed Research and Development on Hand-Held Calculators in School Mathematics makes the following recommendation:

> Recommendation 15. There should be development and research focused on new approaches to achievement testing. Testing policy should be analyzed and appropriately changed in response to these results.
>
> (NIE and NSF, 1976, p. 41)

The report of the Euclid Conference working group on Research Priorities states:

> Basic research on the development of instruments to measure processes as well as achievement in mathematics must be considered. New techniques must be developed and repeatedly used. Research is needed on alternatives to the standardized tests in use today. Our tests are inadequate, but little research is being done to explore acceptable alternatives.
>
> (NIE, Vol. II, 1975, p. 25)

The evaluation chapter of the NACOME report concludes with the following paragraph:

> Evaluation is an essential and useful phase of any instructional program. But to make evaluation play a positive and effective role in school mathematics today there is an urgent need to develop a much broader collection of measurement techniques and instruments and to match these evaluation tools more appropriately to the varied purposes of evaluation.
>
> (NACOME, 1975, p. 135)

The initiative for needed research and development is unlikely to come from the school systems or the state. The technology and expertise in the testing industry are geared to norm-referenced testing and there seems to be considerable inertia which impedes needed change. Consequently, needed developments in testing are likely to be slow in coming unless they are stimulated by the federal government.

Some Factors to Consider in Mathematics Testing

To what degree does testing contribute to mathematics avoidance? Faculty members teaching in the Mathematics Anxiety Program of the University of Minnesota Center for Continuing Education for Women found in interviews with students that they tended to equate their lack of proficiency in mathematics with a lack of intelligence. Furthermore, they did not associate their lack of proficiency at anything else with a lack of intelligence. The students tended to be reasonably well-educated, middle-age women who were good readers. They associated mathematics with intelligence because there was a mathematics component on every intelligence test they had taken. Evidently, experience in taking those tests had contributed to their mathematics anxiety and their avoidance of mathematics. We have no idea of the amount of damage that may have been done to people by testing. We do know that women and minorities tend to shy away from high school and college mathematics courses and in the process they severely limit their opportunities for further education and careers.

Cultural bias in testing is receiving more attention. There is a greater attempt to use test items which fall within the cultural experience of all members of the population being tested. More attention is also being paid to maintaining a balance between males and females, to including a diversity of ethnic names, and to avoiding sex and race stereotypes. The cultural interaction with respect to the reading component of mathematics tests and to the general testing operation including the speeding of tests also needs attention.

A major concern in testing is the question of who decides which tests will be used and what the testing conditions will be. In most cases, those decisions are made at the local level. In some case, decisions are made at the state level as with the New York regents exams and the Florida minimum competency tests. There is strong resistance to a federally controlled national testing program that surfaced at the National Conference on Achievement Testing sponsored by HEW in March, 1978. (Pipho, 1978, P. 586) Though decisons are usually made at the local level, the choices are severely limited when data on comparison with students in other schoool systems is sought. The general situation is that schools have a choice between several standardized tests that have been developed and normed nationally by the test publishers. For the comparison to be valid, local options on conditions of testing are generally not available. For example, local school systems do not have the opportunity to make decisions on reading questions to the students, on allowing use of calculators, or on allowing students all the time they need to finish. Furthermore, with respect to basic mathematical skills, there is wide variation in the rate that school systems are moving; this is particularly true in areas such as use of calculators, and in basic skills like use of the metric system which have only recently received attention. Today, with wide availability of hand-held calculators, speeded computation tests are no longer appropriate (with the exception of tests on basic facts and possibly on ability to make rapid approximations).

The secrecy issue surrounding testing is another item of major concern. We often hear that teaching to the test is a "No, No." The test makers invest considerable money into developing and norming valid and reliable tests so the schools are expected to take careful precautions to see that the tests are not compromised. With increasing pressures for accountability there is more temptation to cheat by persons whose programs are being evaluated. The performance contracting debacle is a prime example. The attempt for secrecy may be a losing cause because, as stated by Anthony Polemeni, Director of the New York City Board of Education Office of Educational Evaluation, "Once a test is administered, it can no longer be considered secure." (Polemeni, 1977, p. 51) Furthermore, there is an inherent inconsistency in the idea of keeping students in the dark about what they are expected to know. Moreover, the secrecy surrounding mathematics testing may be a major factor contributing to mathematics anxiety.

A concern with both norm-referenced and criterion-referenced testing is the potential tendency to concentrate on skills which are easily measurable. For example, testing to determine if students can read tables, charts and graphs is relatively simple. However, it is not as easy to rapidly determine if students can use tables, charts and graphs to organize data. In basic skills testing there is a danger of focusing on isolated low-level skills and neglecting to determine if these skills can be effectively combined to solve problems. This particular testing issue is a potential battleground for advocates of behavioral or cognitive psychology. It is also related to a possible dichotomy between higher education and elementary/secondary education. Post, Ward and Willson found that an overwhelming majority of college and university mathematics educators identify themselves as primarily cognitive in orientation. On the other hand, they observed school teachers, principals, and school practices to be essentially behaviorist in orientation. (Post, Ward and Willson, 1977, p. 338) This apparent divergence may have critical implications for test development, as well as for curriculum and staff development.

The National Science Foundation sponsored Case Studies in Science Education (CSSE) report notes that the major activity of students in mathematics classes at all levels is working assigned problems. (CIRCE and CCC, 1978, p. 19:9) This is perfectly consistent with the way they are evaluated for their work in mathematics. If other types of activity are desired, then more imaginative evaluating techniques appear to be in order.

Mathematics Testing in the Schools

The Case Studies in Science Education Report found dismay on the part of teachers and administrators with the amount of time scheduled by school districts for testing. Furthermore, they found that teachers made virtually no use of standardized test results. (CIRCE and CCC, 1978, pp. 15:21-22) This raises the question as to whether students are subjected to over testing as a result of requirements by school districts, by states and by federally funded programs. School systems should continually monitor their testing programs to see that the information gained is needed and used. They should also minimize the amount of testing by using sampling procedures where possible. Another problem with district-wide testing programs is converting the massive amount of data collected into useful information and providing it to the appropriate decision makers before it becomes ancient history. The CSSE report indicated that the elaborate information systems established in some systems were "cumbersome, expensive, distracting, and apparently of questionable validity for improving the operations of instruction." (CIRCE and CCC, 1978, p. 19:15)

Teachers generally do not use information from standardized tests because the tests usually do not measure student progress toward the instructional goals of the class, because the tests usually do not provide specific diagnostic information, and because too much time usually elapses between the administration of the test and the return of the information to the teachers. In addition to informal evaluation techniques such as observation, teachers most often use tests they construct themselves or publishers' end-of-chapter or unit tests. In some cases, they also use tests prepared by the department, school, district or state. Probably the greatest use of tests by teachers is for grading purposes, with less attention being paid to diagnosis and to item analysis for the purprose of improving instruction. One problem with locally developed tests is that their quality is not always high. A limitation of locally developed tests is that they do not provide information about comparisons between the performance of local and other students. Locally developed tests can provide some comparative information if they use selected items that have been released from the National Assessment of Educational Progress tests.

Programs of instruction which are tightly structured to elaborate testing programs have not generally lived up to initial expectations. Today, little programmed instruction is found in the schools. Performance contracting is dead. Talking typewriters are silent. Self-paced instruction has not produced anticipated achievement gains. For example, Harold L. Schoen reports that the results of 55 subtests in studies comparing achievement in traditional instruction and self-paced instruction (IPI, PLAN, and local programs) in grades K-9, seven favored self-paced instruction, 19 favored traditional instruction, and 29 indicated no significant difference. (Schoen, 1977, p.202) In today's financial environment the greater expense of self-paced instructional programs causes even more caution in adopting programs with elaborate testing components. The problem with these programs may not be so much that they have elaborate testing components as that they place heavy reliance on motivation of individual students and on their ability to learn directly from printed materials. Perhaps after all, learning is a very human endeavor that usually requires a great deal of interaction between the student and the teacher and fellow students. Furthermore, the teacher must be relatively free of record keeping chores in order to be able to teach.

Testing can be used effectively to improve achievement in mathematics. For example, in Minneapolis with the implementation of a mastery learning program of minimum essentials of first-year algebra, achievement in basic algebra skills increased from a 54% level to an 88% level. (Taylor, 1973, p.118) In Phoenix, the implementation of a basic skills program has provided substantial gains. (McCully, 1978, pp. 113-116) Information from testing can also point out areas where resources are needed. For example, in Florida as a result of the statewide competency tests the legislature has allocated an extra $10 million for basic skills instruction this year and they are planning to allocate $26 million next year. Since students fared worse on the math test than on the reading test, the lion's share of these funds is expected to be spent on mathematics instruction. Educational decisions are often made on a crisis orientation and one way to get a decision is to create a crisis.

Generally standardized tests do not directly influence classroom instruction. However, the growing public concern about student achievement, particularly in reading and mathematics, has increased the pressure on school district administrators, building principals and classroom teachers to improve test scores. This pressure is increased even more when standardized test scores for each school are published in the paper. If there is a strong desire to increase those scores, it can be done. For example, in Minneapolis there was a concern that mathematics achievement was several months below national norms at the intermediate level.

An analysis of the standardized test revealed that of 40 computation items, 15 dealt with fractions. After a flurry of curriculum development and inservice on computing with fractions, the test scores rose to grade level according to national norms. A survey of several schools with the greatest increases that the following procedures appeared to contribute to gains:

1. Be aware of objectives tested.
2. Group children and teach to topics.
3. Formulate a good record keeping system.
4. Provide 45 minutes or more of daily instruction.
5. Set aside another time period for maintenance.
6. Present maintenance problems with a mixture of operations.
7. Do daily drill on basic facts.

With these procedures and some practice with multiple-choice questions, the children are much better prepared to take the tests; moreover, teachers report that student anxiety decreases. In fact, a number of the students seem to have the attitude, "Let us at it. We'll tear it apart."

A problem with these procedures is that you can almost never find a standardized test with items that exactly match your objectives. For example, with the advent of the hand-held calculator, the forthcoming change to the metric system, and the difficulty many children have with fractions, there is growing agreement that most of the computation with fractions should be delayed until junior high school. In fact, a conference on fractions sponsored by the Minnesota Council of Teachers of Mathematics and the Minnesota State Department of Education made that recommendation. However, since the standardized test contains a high proportion of fractions and there is pressure to raise test scores, fractions are emphasized in the intermediate grades. It appears that the number of fraction items on the test is high because fraction items are good discriminators and their inclusion contributes to greater reliability of the test. If this is the case, then one is tempted to ask if the current testing technology is the servant or the master of education.

Citywide testing at the junior high school level in Minneapolis utilizes locally developed criterion-referenced tests. Generation rules for each of the items have been written and a new form of the test is generated each time the test is administered. This procedure minimizes the problem of test security and it enables students and teachers to use previous test forms for practice. In effect, the testing program is a means of communicating priority junior high school objectives. Teachers are encouraged to teach to the objectives of the test. The tests are developed and generated by teachers. All mathematics teachers have the opportunity to suggest changes. Periodically, the test objectives are modified as a result of these suggestions. This testing program has the limitations that it does not produce comparisons of local students with national norms and there is no guarantee that different forms are exactly parallel. With respect to the parallelism of forms, there is a tradeoff to be considered when writing the generation rules. If the parameters are narrowly set, then the items generated will be more alike, but in the process the objective will be narrowed.

The idea of generating test items is easily applicable to the computer. Minneapolis teachers have developed a program called COMPUTE which enables teachers and students to use terminals in their schools to generate tests and work sheets (along with answer keys) for approximately 500 different computation objectives. In one school, computation scores on standardized tests increased by three grade levels after the teachers effectively incorporated use of COMPUTE into their mathematics instruction. Another program for generating exercises for about 1,000 different objectives in number properties, geometry, measurement and problem solving has just been completed by Minneapolis and St. Paul teachers and students. Both programs were funded by the Minnesota Council on Quality Education.

The capability to easily generate tests and work sheets from objectives selected by the teacher is particularly useful in a basic skills program because reteaching and retesting are usually involved. Test generation is just one of many potential applications of the computer for testing and instruction. The computer has potential for relieving the teacher from some of the chores of test scoring and record keeping. Creative uses of the computer have particular potential for improving instruction in basic skill areas such as problem solving, estimation, and computer literacy. In spite of the financial plight of the schools, decreasing computer costs and rapid developments in the micro computer field make the increased application of computer technology to instuction appear promising. One of the problems with applications of technology to education in the past has been that it has been touted as a means of replacing "bad" teachers. This has never happened. However, there are many cases where technology has made "good" teachers even more effective.

The annual meeting of the National Council of Teachers of Mathematics held in April, 1978, included a series of sessions on minimum competency in which representatives of various school districts and states reported on the progress and the problems faced in their minimum competency programs. The need surfaced for an effective means of sharing on a national basis information about test development and the implementation of minimum competency programs. There was a repeated admonition of trying to avoid hasty implementation before the testing procedures and instructional programs could be carefully worked out. Many of the attendees at these sessions expressed a need for inservice in test development.

Summary and Conclusion

The major purpose of testing is to provide information to the decision makers at the various levels. The schools today are caught in a squeeze between declining revenues and increasing public pressure for all students to achieve competency in basic skills. A systematic development of cost effective programs of instruction in basic skills is needed.

The major goals of mathematics education fall into three categories: (1) basic skills needed by all citizens; (2) mathematics needed for careers; and (3) affective and cultural aspects of mathematics. The ten basic skills areas set forth in the National Council of Supervisors of Mathematics Position Paper on Basic Mathematical Skills are widely accepted by the mathematics education community.

Standardized norm-referenced testing is inappropriate for individual student assessment and for program assessment, particularly in the area of basic skills. There is an urgent need for research and development in testing and evaluation.

Some factors of concern for mathematics testing include: mathematics anxiety and potential cultural and sex bias; the issue of local control of testing; the issue of secrecy that surrounds standardized testing; and the potential concentration of testing on easily measurable skills.

In the school systems, district-wide information systems which handle test information can become cumbersome, expensive and distracting. Teachers generally do not use standardized test information. Self-paced instructional programs have not lived up to expectations. However, testing can be helpful in improving instruction and identifying needs. Test scores can be increased, but in the process standardized tests can end up dictating the curriculum. Sets of generation rules can be used in the hand generation or computer generation of criterion-referenced tests. The use of computer technology has great potential for testing. There is a need for a means of sharing information about local and state testing programs on the national level.

In light of the factors considered above, there is a clear and urgent need for a national effort of research and development in testing and other means of evaluation in mathematics. There is also a need for a national clearinghouse to facilitate exchange of information about local and state testing programs, particularly in the area of basic skills. The urgency of these needs is precipitated by the rapid implementation of minimum competency programs throughout the country. These national efforts can be accomplished in a manner that will not result in a trend toward federal control of testing. On the contrary, these efforts are needed to free local districts and states from the amount of national control that is inherent in the current reliance on standardized tests. For example, suppose the schools had access to pools of criterion-referenced items or capability for generating those items. Suppose, further, that the schools had information about national performance on these items, such as the information provided by the National Assessment for Educational Progress. Then, the schools could generate tests from locally selected objectives and compare the performance of their students with students throughout the nation on an item-by-item basis. Furthermore, the test generation capability would alleviate the need for secrecy. The research and development efforts should make full utilization of the expertise of school personnel, subject matter specialists, and educational psychologists. Optimal advantage should be taken of new developments in technology, particularly computer technology. As the NSF sponsored Case Studies in Science Education report concluded, the teacher is the key to the quality of instruction that students receive. (CIRCE and CCC, 1978, p. 19:2) Therefore, a major focus of the research and development should be on the classroom--What information does the teacher need? How can testing and other means of evaluation best provide it?

The very fact that the National Institute of Education convened this conference gives rise to optimism that needs in the area of testing will be addressed.

References

Center for Instructional Research and Curriculum Evaluation and Committee on Culture and Cognition. Case Studies in Science Education. Urbana-Champaign: CIRCE and CCC, January, 1978.

McCully, Ron. Basic Skills Program in Phoenix. The Mathematics Teacher. Vol. 71, No.2, February, 1978, pp. 113-116.

National Advisory Committee on Mathematical Education. Overview and Analysis of School Mathematics, Grades K-12. Washington, D.C.: Conference Board of the Mathematical Sciences, 1975.

National Council of Supervisors of Mathematics. Position paper on basic mathematical skills. The Mathematics Teacher, Vol. 71, No. 2, February 1978, pp. 147-152.

National Institute of Education. Conference on Basic Mathematical Skills and Learning, Euclid, Ohio; Volume I: Contributed Position Papers. Washington: NIE, October, 1975.

National Institute of Education. Conference on Basic Mathematical Skills and Learning, Euclid, Ohio; Volume II Working Group Reports. Washington: NIE, October, 1975.

National Institute of Education and National Science Foundation. Report of the Conference on Needed Research and Development on Hand-Held Calculators in School Mathematics. Washington: NIE and NSF, 1976.

Pipho, Chris. Minimum Competency Testing in 1978: A Look at State Standards. Phi Delta Kappan, Vol. 59, No.9, May, 1978, pp. 585-588.

Polemeni, Anthony. The Need for Test Security. Proceedings of the National Conference on Minimum Competencies: Trends and Issues. New York: Center for Advanced Study in Education, The Graduate School and University Center of the City University of New York, 1977.

Post, Thomas R., Ward, William H., Jr., and Willson, Victor L. Teachers', Principals', and University Faculties' Views of Mathematics Learning and Instruction as Measured by a Mathematics Inventory. Journal for Research in Mathematics Education, Vol.8, No. 5, November, 1977, pp. 332-344.

Rotberg, Iris C. and Wolf Alison. Compensatory Education: Some Research Issues. Compensatory Education Staff, Office of Educational Equity, National Institute of Education. November 1974.

Schoen, Harold L. Implications of Research for Instruction in Self-paced Mathematics Classrooms. Organizing for Mathematics Instruction. 1977 Yearbook. National Council of Teachers of Mathematics. Reston, VA: NCTM, 1977.

Taylor, Ross. The Effect of a Mastery Learning of Minimum Essentials Requirement on Achievement in Ninth Grade Algebra. Unpublished dissertation, University of Minnesota, 1973.

Taylor, Ross. The Question of Minimum Competency as Viewed from the Schools. The Mathematics Teacher. Vol. 71, No. 2, February, 1978, pp. 88-93.

INFORMATION PROCESSING ANALYSES OF MATHEMATICAL PROBLEM SOLVING

Joan I. Heller and James G. Greeno
Learning R&D Center
University of Pittsburgh

A growing body of psychological theory has focussed on the processes and structures involved in human cognition. Information-processing analyses have been performed and models of performance developed at increasingly sophisticated levels of specificity. An understanding of general problem solving has been emerging from these efforts, while analyses of problem solving in specific scientific and mathematical domains have contributed a number of additional concepts. The objective of this chapter is to summarize current understanding of the psychology of mathematical problem solving by reviewing these bodies of knowledge and synthesizing concepts from general problem solving theory with those from the technical problem solving literature.

General Problem Solving Theory

Interest in human cognitive processes and representation of knowledge structures can be traced back to early European psychologists such as Selz (1913, 1924) and Bartlett (1932). Gestalt psychologists, including Duncker (1945), Kohler (1927), and Wertheimer (1945/1959), stressed the importance of understanding in achievement of problem solutions. These Gestalt theories emphasized the insightful nature of problem solving. They explained problem solution as involving a sudden understanding of a situation and an integration of previously learned responses in a novel way.

A substantive theory of the ways in which humans process information to arrive at problem solutions emerged with the work of Newell and Simon (1972). They conceptualized problem solution as the successful outcome of search processes, and provided a language, in effect, for expressing and operationalizing concepts central to a cognitive theory of problem solving.

Newell and Simon's theory is based on extensive analyses of adults solving short, well-defined problems of a symbolic nature. Although the three main tasks used -- chess, symbolic logic, and cryptarithmetic -- comprise a narrow sample of problem types, the theory is proposed to account for a broad scope of behavior. Central to Newell and Simon's approach is the assertion that humans, when engaged in problem solving behavior, can be characterized as information-processing systems (IPS's). As a theoretical construct, an IPS is proposed to account for the basic mechanisms of cognition. Some of the strong, central assumptions about the components of cognition include the presence of long-term and short-term memories (LTM and STM), and processes of retrieval and storage in memory, pattern recognition, comparison processes, and symbol manipulation. Explanations of human problem solving behavior at this level of detail have been found to be amenable to computer simulation. A theory of human problem solving behavior can therefore be expressed as a program on another IPS (a computer), and the adequacy of the program for modeling human behavior can be evaluated through comparison of the two systems' performances. This is especially true because an information-processing theory is dynamic -- i.e., implemented as a computer program, the theory describes in detail the changes in a system's knowledge state and sequences of processes through time. Computer traces are typically compared with protocols of human subjects "thinking aloud," resulting in validation or modification of the theory tested.

Newell and Simon constructed programs and tested them against human performance in the manner described above. Because human performance could be adequately accounted for empirically in terms of mechanisms like those in their programs, the resulting description of problem solving processes were considered characteristic of any problem solving IPS, including humans. The characteristics of problem solving here described are those which Newell and Simon assert to be invariant over task and problem solver.

All problem solving is posited to take place in a problem space. Newell and Simon assert that a problem solver must construct and work within an internal representation of the task environment; the problem space contains not only the actual solution but all possible solutions the problem solver might consider. Presented with a particular problem, the individual must encode the components of the problem in a space consisting of the initial as well as desired goal situation, a set of elements representing intermediate knowledge states, a set of operators, or information processes, that may be applied to produce new states of knowledge from existing states, and all other available concepts or relations needed to understand these situations. Thus, this problem space includes the bases for all overt behaviors eventually exhibited by the problem solver as well as behaviors only considered in thinking about the problem. For example, in a chess player's problem space the elements or knowledge states are all possible chess positions; the initial state is the starting board position; the final state is indefinite but is characterized by attainment of a checkmate situation; the operators are legal moves, etc. In cryptarithmetic (a puzzle requiring assignment of distinct digits to letters such that, for example, the statement "DONALD + GERALD = ROBERT," would be arithmetically true, given that D is 5), the elements are all possible combinations of digit-letter assignments; the initial state is the configuration given, with the knowledge that D is 5; the goal situation is not known explicitly but consists of assignment of digits to all letters such that the problem is arithmetically possible; operators are the processes for making all possible assignments; other relations or concepts needed might include parity, equality, and inequality.

Problem solving is postulated to take place by search in a problem space. This statement must be understood in conjunction with the fact that other behaviors relevant to problem solving may be exhibited as well -- i.e., when a problem is presented, it must be recognized and understood, a problem space must be constructed or evoked from LTM, and problem spaces can be changed during the solution process. These aspects of solution are not themselves searches in a problem space; they are crucial contributory activities but final solution is achieved through a search process. Search consists of a consideration of one knowledge state after another until a desired knowledge state is reached -- that is, problem solving is concerned with finding a path through the space from initial problem state to goal state. Movement to new knowledge states is accomplished by analysis of the current problem situation, the selection and execution of operators to produce new states. Search sometimes involves backup -- return to old knowledge states and abandonment of some knowledge state information. This occurs when a dead-end or contradiction is reached (e.g., realization in cryptarithmetic that a required digit-letter assignment is impossible because the digit has already been assigned).

Means-end analysis and goal-directedness are major characteristics of problem solving activity. A problem solver proceeds toward problem solution in a recursive, step-wise fashion, working to select means (operators) to achieve ends (goals). Use of means-end analysis involves comparison of the current knowledge state with the desired goal state to detect differences. Having recognized a difference, the problem solver sets a goal, taking action to remove the difference. An operator or transformation rule is sought to transform the current status. Using a table of connections that indicates which operations are relevant to which differences, the system either finds an operator that accomplishes the desired transformation, or sets a subgoal to reduce a high priority difference in a way that moves the state closer to the goal. Through a series of such transformations, the problem eventually reaches the point where application of an operator removes the remaining difference between existing state and goal state, and the problem is solved.

Planning processes are used to construct proposed solutions in general terms before working on detailed solution steps. Means-end analysis is limited by its mechanism of seeing only one step ahead at any given point. Planning counters this limitation, allowing the problem solver to explore the usefulness of a general strategy for reaching the desired goal state. As described by Newell and Simon, the planning method consists of:

1. Abstracting by omitting certain details of the original objects and operators.

2. Forming the corresponding problem in the abstract problem space.

3. When the abstract problem has been solved, using its solution to provide a plan for solving the original problem.

4. Translating the plan back into the original problem space and executing it. (Newell & Simon, 1972, p. 429).

Thus, the system explores the feasibility of solution plans by identifying general features of the problem situation and goal, and considering the general kinds of changes producible by available operators.

A further analysis of planning in problem solving was given by Sacerdoti (1975). In Sacerdoti's system, called NOAH (for network of action hierarchies), there is knowledge stored about actions that occur at different levels. For example, there is knowledge about the action of moving an object, which is known to include component actions of picking up the object, walking to another location, and setting the object down. For each action at whatever level, NOAH has knowledge of prerequisite conditions that must be present for the action to be performed, and consequences that result from performing the action. Stored knowledge of prerequisites and consequences enables powerful planning procedures, in which a sequence of actions can be planned so the prerequisites needed for future actions are considered in deciding which other actions will be included in earlier parts of the sequence.

The central concepts outlined above have been the subjects of subsequent theoretical and empirical examination. Expansion and further specification of the concepts have resulted and will be discussed in later sections of this chapter.

Newell and Simon's work provided the most thorough, detailed account of human problem solving yet achieved. However, there remained a need for strong theoretical development at the level of general psychological principles that explain performance in broad classes of problems. In an effort toward developing a coherent, general theory of human problem solving, Greeno (1978b) outlined a set of features and processes required in problem solution. His effort to conceptualize problem solving concepts in a general theoretical framework can be utilized to examine the applicability of general problem solving theory to the domain of mathematical problem solving.

Based on hypotheses about the general kinds of psychological skills required for problem solution, Greeno suggested a typology of problems comprised of three ideal types. Problems of inducing structure were described as those for which some elements are given and the task requires identification of the pattern of relations among the elements (e.g., analogies, series extrapolation). Greeno proposed that the main cognitive ability required for solving problems of this type is a form of understanding, i.e., apprehending relations and constructing an integrated representation. Problems of transformation involve an initial situation, a goal, and a set of operations for producing changes in situations (i.e., move problems such as Tower of Hanoi, and change problems including theorem proofs). Solution of transformation problems relies primarily upon use of means-end analysis and planning processes. Tasks referred to as arrangement problems are those that present some elements and require the problem solver to arrange them in a way that satisfies some criterion (e.g., anagrams, cryptarithmetic). Solution requires a process of constructive search, requiring generation of the possible knowledge states that constitute the search space, and search for the solution in that space. Greeno asserts that most problems are not pure cases of one problem type, but rather include strong components of the three types and involve corresponding demands for the various problem solving processes.

The remainder of this chapter will include discussion of the knowledge and processes required for solving problems in mathematics, and for solving practical problems in which mathematical knowledge is required. The latter discussion will focus on solution of word problems, where studies of the processes have been conducted. Our goal in this discussion is to use the general concepts that have been developed in the theory of problem solving to provide an integrative framework for discussion of the available literature on solution of technical problems that are relevant to mathematics instruction. First, we will discuss problem solving processes that are required for solving ordinary exercises of the kind that are used as homework and test problems. The final section will present analyses of problem solving involving applications of mathematical knowledge.

Mathematical Problem Solving

Most mathematical instruction consists of training for solving specific kinds of problems, and the activity that students engage in as they do exercises or take tests consists of problem solving processes. Although most mathematics educators consider "problem solving" to refer only to relatively unusual parts of the curriculum (thus feeling that there should be more attention to problem solving in the curriculum), this view does not take into account what it is that students must actually do in order to succeed in the ordinary tasks that constitute most of the curriculum.

The claim that ordinary exercises require problem solving is consistent with many persons' intuitions about geometry. Consider the problem: "Find the length of a diagonal of a rectangle whose perimeter is 20 in. and whose width is 4 in.," (Jurgenson, Donnelly, & Dolciani, 1972, p. 323). This has the structure of a transformation problem. There is an initial situation, consisting of given information, and there is a goal of finding a quantity that is not provided but that can be inferred from the given quantities. Problem solving operators involve rules for inferring quantities from other quantities, for example, inferring the length of a rectangle from its perimeter and width. The problem is solved when a sequence of operations leads to an inference that assigns a quantity to the diagonal. Another example is the problem, "Prove: If the diagonals of a parallelogram are congruent, the parallelogram is a rectangle," (Jurgenson, Donnelly, & Dolciani, 1972, p. 243). This problem has an initial situation consisting of some statements, with the goal of forming a proof of one of the statements. This is an arrangement problem, where the student must construct a sequence of statements that are related to each other according to the rules of proof. The structure that is to be produced starts with given information and consists of a series of inferences that terminates with the statement to be proven. Thus, standard proof problems in geometry have components of both arrangement and transformation problems, since the structure that students must arrange is a series of transformations.

Standard exercises in algebra also have the defining features of problem situations. The following is an example: "Solve the following compound sentences: $5x + y = 15$ and $3x + 2y = 9$," (Payne, Zamboni, & Lankford, 1969, p. 230). The initial situation is the pair of equations that is given. The goal is a different pair of equations that are to be derived, assigning values to x and y. Each step in the solution involves a transformation of an equation that is given or that has been derived, or a derivation of a new equation from two earlier equations. Another example is the problem of simplifying an expression, such as (Payne, Zamboni, & Lankford, 1969, p. 394):

$$\frac{x^2 + 8x + 16}{x^2 - 9} \cdot \frac{x - 3}{x + 4}$$

Problems of simplifying expressions are also problems of transformation, but the goal is not well defined. The given expression is the initial situation. Problem solving operators are the rules for factoring, cancelling, and so on. The problem is solved when an expression is found for which no further problem solving operators can be applied.

The claim that mathematical exercises require procedures of the kind needed for problem solving also holds for standard computational problems. Although some readers may be unconvinced by the assertion that even a simple computational exercise presents an initial situation and a goal that must be reached by a sequence of operations, it is harder to remain skeptical after seeing the results of a serious attempt to analyze the procedural knowledge required for a system that performs simple column-addition and column-subtraction problems, much less the knowledge required for more complex operations such as those involved in fractions. Consider subtraction. The problem solver must know how to break the problem into subproblems, each with its subgoal of finding the correct number to place in each column of the answer. In each subproblem, a condition must be tested to determine whether the subtraction operation (top minus bottom number) can be performed. If it cannot (because the top number is smaller) an additional subgoal must be formed that is satisfied when the situation is transformed (by borrowing) so the subtraction operation can be performed. An analysis complete enough to provide for typical errors has been given by Brown and Burton (1977). The knowledge structure they hypothesized for the task of subtraction is shown in Figure 1. Its structure is that of a procedural network, a formalism developed by Sacerdoti (1975) to represent knowledge for problem solving somewhat more complex than that generally represented by production systems. While it is unusual to include "simple" computational exercises in the domain of "problem solving," it seems to us that this does a considerable disservice to young students who acquire very complex knowledge structures in order to perform these tasks, especially when analysis of their knowledge reveals that its structural features are as complex as any that have been developed in the theory of problem solving.

A preliminary analysis of strategies in solving algebraic equations has been given by Bundy (1975), and discussed by Brown, Collins, and Harris (1977). In Bundy's analysis, knowledge for solving equations includes a basic method, consisting of three strategies, and an auxiliary method for removing nasty function symbols. The strategies in the basic method are called isolation, collection, and attraction. Isolation involves operating on an expression so that the unknown term is not embedded in expressions for functions. Collection involves reducing the number of occurrences of the unknown variable. Attraction involves transformations that bring

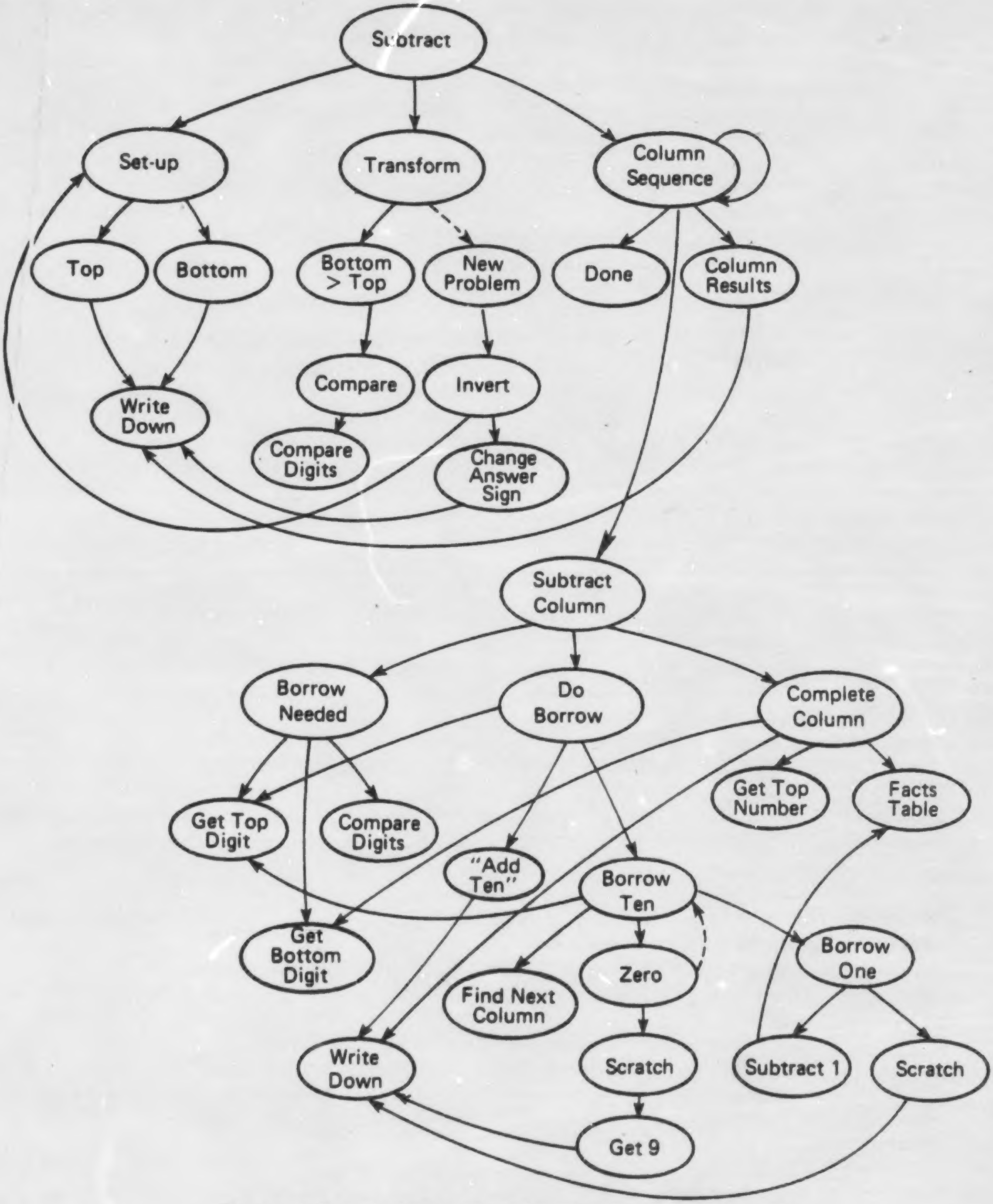

Subtract
Set-up
Transform
Column Sequence
Top
Bottom
Bottom > Top
New Problem
Done
Column Results
Compare
Invert
Write Down
Compare Digits
Change Answer Sign
Subtract Column
Borrow Needed
Do Borrow
Complete Column
Get Top Number
Facts Table
Get Top Digit
Compare Digits
"Add Ten"
Borrow Ten
Get Bottom Digit
Borrow One
Zero
Find Next Column
Scratch
Subtract 1
Scratch
Write Down
Get 9

Figure 1. A Procedural Nework for Subtraction (Brown & Burton, 1977)

occurrences of the unknown term closer in the expression; for example, bringing two occurrences to the same side of the equation. Strategies for removing undesired function symbols include inversion, such as squaring both sides of an equation to remove a radical, and use of definitions. These strategies correspond to a set of subgoals that the problem solver adopts in order to make progress on the problem. Instruction in algebra generally focuses on the operations that are permitted in transforming expressions and equations. Bundy's analysis makes it clear that knowledge of algebraic transformations is not sufficient for solving equations. The student must also have strategies and methods for choosing operations that will be helpful in making progress toward a solution. These strategies and methods are of the same general kind as the strategic knowledge that is critical in any problem solving system.

An analysis of processes used in solving various geometry problems has been developed (Greeno, 1976, 1977, 1978a). A model, called Perdix, simulates the process of proving theorems and computing the measures of specified angles or segments given measures of other things in a diagram. These are the standard exercises of geometry. It is clear that the processes required to successfully complete these exercises are the kind that we characterize as problem solving processes. They include strategic knowledge for setting subgoals and planning, as well as knowledge for recognizing patterns and making inferences based on general propositions.

Some results of the analysis of geometry problem solving have provided relevant information for a theoretical question that has been of interest in the psychology of problem solving. The question involves processes involved in solving ill-structured problems, and whether principles that apply to solving well-structured problems also apply in situations where problems are ill-structured. This question arose because much of the theory of problem solving has been developed by analyzing problems in which the initial situation, the goal, and the permissable problem solving operators are all definitely specified. Since many problem situations lack specificity of one or more of these components, generality of analyses such as Newell and Simon's (1972) has been questioned, for example, by Reitman (1965).

Two aspects of ill-structured problems appear in ordinary geometry exercises. One aspect involves indefinite goals. When triangles are proven congruent, the pattern that is found (SSS, SAS, ASA, or whatever) is not specified in advance, nor are specific patterns usually considered and used as targets for search (Greeno, 1976). The second aspect involves incomplete specification of the problem space by the initial situation of the problem. Problems that require construction of an auxiliary line present less than a complete initial problem space, and the problem solver must enrich the initially presented material in order to find a solution (Greeno, 1978a).

The processes needed to solve problems with indefinite goals or constructions do go beyond those developed in the theory of well-structured problem solving. However, they seem quite compatible with the processes of means-end analysis and planning developed in that theory, and should be considered as extensions of the earlier theory, rather than as fundamental revisions of it. To account for use of indefinite goals we can postulate that problem solvers represent goals using a pattern recognition system, rather than as a single object or combination of features. A pattern recognition system is able to identify any of a set of alternative patterns, and thus can be used to determine whether any of several possible ways of achieving a goal has been reached (Greeno, 1976). To account for the occurrence of constructions, we can postulate planning knowledge of the kind analyzed by Sacerdoti (1975), in which the individual knows prerequisites for solving problems in certain general ways. For example, one plan for proving that angles are congruent is to consider triangles that have those angles as parts, and then prove that the triangles are congruent in a way that makes the angles corresponding parts of the triangles. The main prerequisite condition of this plan is the existence of triangles that contain the angles. If this prerequisite is not present, it can lead to a construction that produces the needed triangles by adding an auxiliary line to the diagram.

Word Problem Solving

Word problems have been the subject of a number of information-processing analyses. These problems consist of narrative passages describing a situation in which quantities and quantitative relations are important. They require the solver to read and understand the passage, and select and apply mathematical operators or scientific principles to determine the value of one or more unknown quantities. Word problem contexts range from basic, elementary level arithmetic to situations involving high level algebraic relations or technical domains such as statics or thermodynamics in physics.

The major difference between solution of ordinary mathematical exercises and solution of word problems is that word problems require a process of understanding the situation described in narrative passages. The understanding process has been the subject of considerable investigation in recent studies, including Anderson (1976), Norman and Rumelhart (1975), and Schank and Abelson (1977). These analyses have considered the process by which an understander constructs a representation of the information contained in text, including the relationships between concepts that are mentioned in the message. An important factor in understanding is the degree to which the understander fills in missing information based on the general knowledge that the understander has. Depending on the degree of inferential processing carried out by the understander, the representation that is achieved may be relatively superficial, with loose connections among the concepts in the text, or relatively coherent and tightly integrated, with a strong relational structure composed of implications that are drawn from the understander's knowledge.

The possibility of understanding situations in ways that differ in depth and coherence implies that in solving word problems there is an important trade-off between the amount of processing accomplished in the understanding of the problem and the amount of processing left for the problem solving operations. An individual who represents the problem situation in a relatively complete way might be expected to have information in that representation that is useful in guiding the selection of problem solving operators. A weaker representation would provide a situation in which more search and planning would be required.

We will review the major information-processing analyses of word problem solving, and then attempt to synthesize the findings in terms of general concepts of processing skills and knowledge organization across problem domains. A major conclusion of this review will be that skill in solving word problems depends strongly on the problem solver's ability to represent the problem situation in a way that facilitates the processes of search and planning for problem solving.

Algebra word problem solving. Paige and Simon (1966) analyzed algebra word problem solution by comparing the behavior of adult solvers with the strategy of Bobrow's (1968) STUDENT program. STUDENT was an artificial intelligence effort intended to demonstrate the capability of the computer to interpret natural language input and derive mathematical equations in algebra word problem solution. The program solves most problems by a direct translation process -- the problem statement is interpreted phrase-by-phrase, using syntactic function tagging, substitution, and transformation rules to construct a mathematical expression representing the problem situation. For example, a statement such as "the number of customers Tom gets is twice the square of two tenths times the number of advertisements he runs" is transformed directly into an expression, e.g., "$x_1 = 2 * (.2 * x_2)^2$." This is accomplished by assigning a variable, "x_1" to the initial noun phrase, "the number of customers Tom gets;" interpreting "is" as "="; "twice" as "2 *"; "the square of" determines the exponent, 2; etc. While Paige and Simon found this direct translation strategy to correspond with much of their human solvers' behavior, they also discovered evidence for a second solution mode. A number of their subjects constructed "auxiliary representations" of problem situations, generally in the form of drawings representing the physical problem situations encountered. Those individuals who typically relied on semantic, substantive information in the solution process were considerably more successful at discovering problem incongruities than subjects who relied on direct syntactic problem interpretation. That is, application of direct translation to certain "contradictory" problems in Paige and Simon's task set resulted in impossible solutions (such as negative values for the size of physical objects), or in transformation of the problem to a related, physically possible, situation. Individuals who realized that these problems were not solvable were generally those who included auxiliary cues from their knowledge of the world in the internal or external problem representations that they constructed. Paige and Simon concluded that problem solvers differ in their reliance upon syntactic versus semantic processing in problem solution, but that most solvers included some amount of semantic information in their solutions. Therefore, they suggested that a theory of skilled algebra word problem solution would have to incorporate the use of semantic knowledge in the problem understanding process to adequately account for human problem solving behavior.

Hinsley, Hayes, and Simon (1977) also distinguished between two modes of algebra word problem solution. They designated as a "text grammar" approach one that emphasizes the importance of formal structure and postulates distributed decision processes. The direct translation process is considered in this category because it involves no choice at the time of reading and makes little use of semantics. The second, "schema," approach emphasizes the importance of semantic knowledge and major decisions occurring early in the comprehension process. STUDENT incorporates this type of process involving choice of schema at the time of reading with respect to "age" problems only. If a problem is an "age" problem, as determined by the presence of particular phrases such as "years old" and "as old as," the problem is solved by inclusion of special heuristics. Hinsley, Hayes, and Simon postulated that, if this second approach were characteristic of human solution of algebra problems, then the availability and reliance upon problem schemata, or problem category information, would be empirically verifiable. In a series of experiments, they did establish that people can categorize problems into types, and that people have knowledge about each problem type that is used in formulating problems of that type for solution. Furthermore, Hinsley, Hayes, and Simon again found evidence of two different solution procedures used by subjects in extracting appropriate equations from problem texts. One procedure was characterized by reading of the entire problem before formulating any equations or noting any relationships explicitly. These solutions included use of problem category information and retrieval of schema that aided problem solution. The second procedure was the line-by-line, direct translation approach described previously. Some semantic knowledge was found to be incorporated in a portion of the solutions of this type, however. Hinsley, Hayes, and Simon's results supported the interpretation that the line-by-line procedure is a default process that is used if the problem is not successfully matched to one of the solver's available problem type schemata. That is, schemata may be important only in the formulation of problems in which the semantics of the cover story match the problem structure in an expected way, and the schema in question is currently available to the solver. This ability to perceive the underlying mathematical structures of word problems and to relate problems on the basis of the perceptions has been investigated by Krutetskii (1976) and Silver (1978). Both investigators found that there is a significant relationship between problem solving ability and the capacity to perceive accurately the formal structure of word problems. In fact, the tendency to sort problems according to mathematical structure was significantly positively correlated with numerous general verbal and mathematical ability measures. Conversely, students of lower ability were found to attend more to contextual and syntactic problem features than to underlying mathematical structure.

Technical word problem solving. Studies of solution processes in technical, scientific problem domains have focussed both on individual differences and the basic nature of processing.

After analyzing the physics problem solving behavior of one novice and one expert subject, Simon and Simon (1978) concluded that the expert's solution generally involved translation of the problem passages into physical representations, then use of those representations to select and instantiate the appropriate equations. They refer to this schema construction as "physical intuition" and contrast it with solving "simply by plugging in the formulas." The novice's behavior was typically "algebraic" in that she appeared to go directly from the problem statements to the equations required to solve them, with no evidence of any mediating cognitive representation. The novice's behavior was further characterized by explicit use of means-end analysis. She typically identified the goal and worked backward, setting up equations to solve for subgoals until the final solution could be reached. Only on particularly difficult problems did the expert rely on this type of problem solving strategy.

Larkin (1977) analyzed the protocols of physics problem solvers working on rather complex problems. She found evidence for an initial "qualitative analysis" performed before any equations were generated. Protocol data revealed that in the preliminary phase of problem solution, expert subjects constructed representations of the physical situation described in the problem. These representations were qualitatively elaborated by inclusion of supplementary information required for understanding the problem situation that was not included explicitly in the written statement.

Skilled solvers were found to retrieve from memory tentative "chunks" or "schemata" of associated physics principles for consideration, and to choose an aspect of the representation to which the chunk might be applied. Problem features are elaborated further if necessary in relation to the chunk being considered; the solver attempts to determine whether the tentatively chosen cluster of principles is applicable to the problem representation. When the match is bad, the chunk or problem aspect is discarded and a new combination considered. Upon finding a matching chunk and representation, the solver seeks a solution procedure. Larkin reported that expert solvers appear to have solution procedures associated with schemata, and that these procedures are prioritized with respect to their utility for solution of problems of different types. Some solvers were observed to evoke equations immediately upon recognizing a schema-problem match, apparently having spontaneously retrieved one highly appropriate solution procedure. Other individuals executed preliminary traces of several solution procedures in general terms to select the most applicable and promising procedure. Once a promising procedure was selected, all of these solvers evoked and instantiated equations to reach accurate solutions. Some solvers, however, executed no preliminary traces, reached dead-ends, and needed to try other procedures until one finally led to solution. These lower skill individuals, that is, skipped directly to equation building and manipulation without an intervening step of checking the usefulness of procedures in a abstract planning space. Alternatively, this behavior can be interpreted as indicating inefficient or no prioritization of solution procedures.

Thus, skilled problem solving apparently involves construction of a comprehensive problem representation as well as evocation of an appropriate theoretical schema or cluster of relevant physical principles for application. Solution procedures are associated with these chunks of knowledge, and instantiation of equations is executed with great efficiency. Evidence of storage of physical principles in related configurations was found only for experts; knowledge appears to be less well organized, and solution procedures less often associated with available schemata, in lower skill solvers.

Skilled physics problem solvers have been seen to be characterized by their translation of the verbal problem statement into an integrated, internal or external representation of the physical situation. This representation is elaborated by inclusion of auxiliary information necessary for understanding the problem, but not included in the natural language passage. Novak (1976) has demonstrated this translation process in a computer program, called ISAAC, that is able to read, understand, draw pictures of, and solve a set of physics problems stated in English. Problem statements are read, parsed, and interpreted semantically to construct an initial internal model of the objects and relations in the problem. Once the problem has been translated in this way, the solution is guided entirely by the internal representation, with no further reference to the problem statement. The initial model is interpreted in terms of canonical object types. That is, based on the system's knowledge of physics, objects in the problem are considered with regard to their functions in the problem -- i.e., a person may be represented as a member of the canonical class of "pivot" when carrying a plank, or as a "point mass" when standing on one. This selection and completion of canonical object frames is considered by Novak to be one of the most important aspects of skilled physics problem solving ability. Once this interpretation has been accomplished, ISAAC constructs a geometric model of the problem, considering the location and orientation of objects in relation to one another. Although ISAAC proceeds to generate and solve equations at this point, it should be noted that human problem solvers would probably be observed to draw external representations of the internal model thus derived. ISAAC has the capacity to draw diagrams based on the model, and does so later in its execution. The major point, though, is that problem solution is critically dependent upon construction of an integrated, elaborated representation of the original problem statement. Since the way in which the problem is understood or represented so seriously impacts subsequent solution, it may be suggested that skilled solvers probably represent problems in ways that facilitate their solution -- that is, the problem space generated itself suggests promising subsequent action. In particular, how and what information is extracted from the problem statement, how it is organized and elaborated with auxiliary information, and how irrelevant information is identified and excluded, all would affect solution processing. In this sense, it may be that skilled problem solvers incorporate planning processes in the initial construction of the problem representation. The relationship between planning and problem representation is as yet unclear, however, and constitutes an interesting issue for investigation.

The initial qualitative analysis phase described by Larkin clearly does include a type of planning required for efficient technicl problem solving. Execution of preliminary traces to identify the most promising solution procedure apparently is performed only by fairly sophisticated problem solvers. Less skillled individuals seem to plunge into the details of equation manipulation without first eliminating possbily unproductive solution paths. Sacerdoti's (1975) computer program, NOAH, embodies a set of mechanisms that accomplish this sort of planning. NOAH is a hierarchical planning system that develops and expands plans to increasingly detailed level of specificity. Crucial to the emergence of an effective final plan is the functioning of "constructive critics" that examine the most developed plan currently under consideration. Each potential action in a plan has associated with it the preconditions, purposes, and consequences of that action. In the plan development process, the interaction of the planned actions is examined (criticized) in terms of their associated implications, and conflicts, redundancies, or inconsistencies are eliminated by reordering, addition of constraints, or elimination of redundant operations. Thus, a self-regulating, critical mechanism governs the construction of purposeful plans. In terms of technical problem solving, it may be said that the skilled problem solvers consider the requisite conditions along with the outcome or consequences of applying particular solution procedures in terms of the purpose or goal pursued (i.e., unknown variables for which quantities are sought). An on-going process of criticism is probably functioning to insure that the procedure(s) selected interact in a way that allows the problem goal to be reached. The development of this critical capacity is a crucial area for further investigation in understanding the nature of skilled problem solving.

Arithmetic word problem solving. Perhaps because of the difficulties involved in obtaining data from young children that allows for process analysis, little work has been conducted so far in this area. We have begun a project to study and model the process of solving arithmetic story problems, and can report some preliminary findings here. In this work we are collaborating with Mary Riley.

As described earlier in this chapter, a model of solving word problems was developed by Bobrow (1968); we have designed a computer simulation that models quite a different strategy of skilled word problem solution (Heller & Greeno, 1978). Bobrow's system relied almost exclusively on syntactic information to translate problem texts into simultaneous algebraic equations for solution. Our system models semantic processing as the main component of problem understanding. The system abstracts the underlying quantitative structure, or schema, from problem passages, using a specialized method of reading comprehension. That is, the system constructs a semantic network representing the information in the problem in a way that explicitly facilitates problem solution. Arithmetic operations are selected on the basis of the structural representations constructed; there is no intervening process of building equations before the operation is selected.

The set of schemata that represent alternative structures of quantitative information were derived from analysis of addition and subtraction word problems found on standardized tests and in elementary school curricula. We have identified three distinct schema that appear to be necessary and sufficient for representing the structures of all one-step addition and subtraction problems we examined. These schemata can be referred to as Change, Combination, and Comparison.

Change situations are those in which some event changes the value of a quantity. For example, statements like "Jan had three lamb chops; Joe gave her four more," demonstrate a change in the quantity of objects possessed by one person as a result of an action. The abstract schema for this type of situation contains three main components: an initial quantitative state, an action involving a change value, either in the direction of increase or decrease, and a resulting quantitative state. The direction of the change, as well as the quantity that is unknown (initial, change, or result), govern the mathematical operation needed.

Combination situations are those in which there are two separate amounts that comprise a third, combined value. For example, in the problem, "Ray saw six birds. Jill saw four birds. How many birds did they see altogether?" the two separate quantities must be combined to determine the value of the unknown total quantity. In Combine problems, the choice of an operation depends on whether the unknown value is one of the separate amounts or the combined amount.

The third schema involves two amounts that are compared and the difference between them. Examples of Comparison problems include, "Sue rode eight miles. Sam rode four miles. How many more miles did Sam ride than Sue?" or, "Rose ate nine cookies. Jim ate three fewer cookies than Rose. How many cookies did Jim eat?" The operation to be performed depends on the direction of the difference (more or fewer), and the quantity that is unknown (the difference or one of the separate quantities).

In our problem solving model, the problem text is translated first into a parsed form, in Anderson's ACT formalism (Anderson, 1976). As each problem proposition is read, it is incorporated into a gradually constructed abstract structure of one of the three types described above. This is accomplished by a translation process somewhat akin to Novak's (1976) canonical object frame construction. That is, objects and actions are considered in terms of their functions in the problem situation. Thus, the entire problem context is considered, but a very abstract representation of only the quantitative relations is constructed based on categorical information stored about verbs, types of phrases, parts of speech, etc., in the system. Details of particular names, verbs, and locations, for instance, are dropped as irrelevant to the schema. The crucial quantitative relations are inferred from the semantics of the problem while inferences that would be made in other contexts, such as ordinary story comprehension (e.g., Schank & Abelson, 1977), do not occur.

When a semantic representation has been constructed, the system infers the appropriate arithmetic solution procedure. It does this by referring to stored information about which operations are associated with which schemata, considering the direction of change or difference and the location of unknown quantities in the representation. At this point, our understanding of the details of this process is largely speculative and undergoing revision as empirical evidence is collected. We will therefore reserve a description of the process for later reports.

The most notable aspects of our theory derive more from its global solution method than the details of its operation. We are postulating that word problems can be solved without generating equations, and that construction of internal representations of the underlying quantitative relations to guide the solution are foremost in effective problem solving. These notions are supported by data showing that children can solve word problems before they begin to learn arithmetic and have any knowledge of equations (Buckingham & Maclatchy, 1930). Our own research (Riley & Greeno, 1978) confirms this finding, with the addition of evidence that the semantic schemata involved in problems are rather strong determiners of problem difficulty for these children. In one experiment, second-grade children were found to have little difficulty with any problems with the Change structure or Combination problems with the combined amount unknown. Students were not as successful with Combination problems with one of the separate amounts unknown, and all of the problems having Comparison structures were relatively difficult for these young children. These results strongly support the notion that construction of accurate problem representations is of great importance in attainment of correct solutions. This process relates strongly to recent developments in the theory of natural language understanding (Schank, 1972; Winograd, 1972). "Understanding" is considered by these theorists as a constructive process: a situation is said to be understood when a complete and coherent structure representing the objects, concepts, and relations therein is constructed. Greeno (1977) has proposed that integration of these ideas with theories of procedural knowledge allow for more complete conceptualization of problem solving with understanding. In fact, much of the most recent work in problem solving described in this chapter has attempted to use this approach, and a general theory of skilled problem solving performance is thereby emerging. More specific distinguishing features of problem solving at different levels of expertise have also been discussed. In the following section we will attempt to synthesize these findings into a coherent picture of word problem solving processing.

Processing in Word Problem Solution with Understanding

Word problems can be grossly characterized as well-structured problems of transformation; definite initial and goal states are both given, and there exists a set of possible operators for moving toward solution. However, mediating representational and initial elaboration processes are required to transform word problem texts into problem spaces suitable for solution with understanding. A procedure for apprehending relations among components allows construction of integrated problem representations in the form of relational structures. Problem solving then requires search for the proper set and sequence of operators (arithmetic operations, algebraic manipulations, scientific principles and/or equations) leading to the final goal state. Equations are evoked or constructed and instantiated on the basis of these analytical processes.

It is becoming clear that the nature of problem solving processing differs as a function of both the expertise of the solver and the difficulty of the problem. That is, the nature of problem solving processes required cannot be determined in isolation from the particular problem solver engaged in working toward its solution; the problem solving strategies of an individual appear to vary depending on the complexity of the task, and different individuals working on identical problems differ in their behaviors. This view of problem solving can be extended to the domain of simple arithmetic; we have proposed that even basic arithmetic operations require complex problem solving processes, especially for beginning students. Although this is not a new concept, the specific forms of problem solving, given different combinations of solver skill and problem difficulty, are just beginning to be specified at a detailed level of analysis. With respect to word problem solution, the literature described in this chapter begins to suggest a general picture of these variations in problem solving behavior.

Problem solving skill may be viewed in terms of the solver's increasing capacity or competence for handling problems. Although numerous studies have undertaken to distinguish between "expert" and "novice" performance in various problem solving domains (e.g., deGroot, 1966 and Chase & Simon, 1973, in chess; Simon & Simon, 1978, and Larkin, 1977 in physics word problem solution), the findings of such analyses are best used for specifying a developmental continuum, i.e., a model of stages or levels in problem solving ability rather than only dichotomous distinctions between high and low skilled performance. The major dimensions along which these differential abilities vary in word problem solution are: the "understanding" processes, i.e., construction of problem representations; the organization of relevant knowledge, including problem category information or presence of chunks of scientific principles; the nature of initial analysis and planning processes; and the overall type of solution strategy used. A chart summarizing the solution features at different levels of competence is provided in Figure 2. Where evidence exists in the literature regarding the solutions described, a citation is provided in the rightmost column. The remaining descriptions are provided as intuitive propositions that remain to be tested empirically.

Solution characteristics are indicated separately for non-technical (arithmetic, algebra) and technical (scientific) word problems. This is to some degree an artificial distinction, since these problems tend to overlap depending on the extent of auxiliary information or principles required for solution. However, the distinctions are useful for examining behavior on clearly different classes of problems.

The continua along which solution features vary are abstracted in Figure 3. For the purposes of this general discussion, some of the details in Figure 2 will be dropped and the more global variations emphasized.

PROBLEM TYPE	PROBLEM REPRESENTATION	KNOWLEDGE STRUCTURE	INITIAL ANALYSIS	SOLUTION STRATEGY	SOURCE
Non-technical	Rely on verbal problem statement (even if diagram constructed); no use of semantic and auxiliary information.	No problem category information used. (Can't tell if available.)	Read problem.	Direct, phrase-by-phrase translation from written problem to equation.	Paige & Simon, 1966 Hinsley, Hayes, & Simon, 1977
Technical		No evidence of schemata; principles available as discrete bits, no qualitative understanding.	No informal qualitative analysis; identify unknown and given values. Go directly to equations.	Nearly random search for equations containing desired quantities; "plug in values" solving for subgoals until answer is isolated.	Simon & Simon, 1978
Non-technical	Some reliance on mediating representation and semantic information; reference to verbal problem statement as needed.	Some category information available; solution procedures not strongly associated with category information.	May recognize problem structure; search for solution procedure not strongly guided by category understanding.	Formally pursue several unproductive solution paths before problem solved; alternate between equation building and problem analysis.	
Technical		Some principles organized into schemata; priorities not well defined.	Partial qualitative elaboration of problem features in relation to available schemata; select tentative procedure without efficient traces to eliminate ineffective plans.	Implement several plans for solution before problem solved; frequent iteration between analysis and equation working; inefficient solution.	Larkin, 1977
Non-technical	Construct diagram when useful; develop integrated internal representation through qualitative analysis; rely only on mediating representation.	Strong problem category information available with associated solution procedures.	Recognize underlying problem structure; search for solution procedure associated with category.	Construct equations suggested by procedure associated with problem type.	
Technical		Principles in well organized prioritized schemata with associated solution procedures.	Qualitative elaboration of problem features in relation to available schemata; search for matching schemata; preliminary traces to determine most promising solution procedure.	Implement plan for solution using selected schema; generate equations; rarely need to rework by alternative procedure.	Larkin, 1977
Non-technical	Diagram often not needed; rely on strong, spontaneous mediating representation.	Strong problem category information available with associated solution procedures.	Problem structure immediately apparent; solution procedure evoked spontaneously.	Immediately generate and solve equations, solving for subgoals if necessary without hesitation.	Paige & Simon, 1966 Hinsely, Hayes, & Simon, 1977 Riley & Greeno, 1978
Technical		Principles in well organized prioritized schemata with associated solution procedures.	One appropriate schema immediately apparent on basis of qualitative problem features.	Immediately apply principles to generate and combine equations.	Bhaskar & Simon, 1977 Simon & Simon, 1978

Figure 2. Features of word problem solution at varying levels of competence.

	PROBLEM REPRESENTATION	KNOWLEDGE STRUCTURE	INITIAL ANALYSIS	SOLUTION STRATEGY
LOW SKILL	Rely on verbal problem statement; syntactic processing	No evidence of relevant schemata; solution information not associated with problem category knowledge	Read problem only	Piecemeal, direct translation; "plug in values"; search for solution combined with equation manipulation
	↕	↕	↕	↕
HIGH SKILL	Rely entirely on integrated mediating representation; semantic processing with qualitative elaboration; problem "understood"	Strong available schemata; solution procedures associated with schemata and prioritized for application	Use information associated with schemata to identify solution procedures; most promising procedure either immediately apparent or determined by preliminary traces	Highly organized evokation of equations based on preliminary analysis and strong structural representation

Figure 3. Continua of word problem solution variation.

The first feature, problem representation, has been discussed previously in terms of the notion of "understanding." Increasing levels of competence are characterized by greater reliance upon an integrated, qualitatively elaborated mediating representation of the problem structure. Syntactic processing in low level solutions, with heavy reliance upon the verbal problem formulation, contrasts with the semantic processing typical of individuals capable of problem solution with understanding. The integration and elaboration of structural problem representations with auxiliary and semantic knowledge is present in greater degrees at increasing levels of competence. Use of external or diagrammatic problem representations often characterizes higher skill solvers, but the critical factor is the reliance upon an internal understanding of the problem's underlying structure.

The organization of relevant knowledge or availability of schemata also becomes more salient with increasing competence. The strength of association between schemata and solution procedures increases as well, with prioritization of procedures more evident in higher ability problem solvers. That is, not only does the higher skill individual have available a broader range of schemata, but solution procedures are associated with the schemata and the procedures are ordered according to their usefulness with respect to different problem types and situations.

Initial problem analysis is a feature strongly related to problem representation, in that effective construction of representations reflects skill in and attention to preliminary elaboration of problems. In contrast to the low skilled individuals' rapid plunge into equation or number manipulation, higher skill individuals tend to consider a problem in an abstract problem space, planning a solution before working with the details. A sort of means-end analysis is sometimes performed, and potential solution procedures are examined before one is chosen as most promising for reaching the problem goal. This search for a solution procedure is highly organized with greater competence, as opposed to the low ability solvers' nearly random search for the combination of operators or principles that might yield the desired value. Lower ability solvers are not guided by an understanding of the problem and therefore cannot develop coherent solution plans with clear goal structures.

It is interesting to note that the initial analysis phase is completed so rapidly in the highest level of competence that it was not explicitly noted in the literature. Although there was no overt evidence for this process in protocols of either very high skilled or low skilled solvers, we suggest that the surrounding behaviors imply considerably different explanations for the absences. In the case of lower skill solvers, the piecemeal, disorganized solutions suggest that no planning processes were executed. In the rapidly executed, high skilled solutions, however, the immediate generation of consistently correct and efficient solutions supports the notion that only one procedure was seen as applicable. In effect, the problem was so routine and simple that more than one procedure did not have to be considered -- the solution was directly and singularly associated with the problem representation constructed. It cannot be concluded that the solution was obvious to the lower skilled individuals.

The overall solution strategy used varies in the degree to which planning and integrated structural representations determine equation construction. Lower skill solutions typically embody search processes in combination with equation manipulation; the solver is seeking a solution procedure at the level of instantiating equations; there is no reliance upon a clear understanding of the problem. In higher skill solutions, the search process is completed separately and before equations are constructed. Solution procedures are considered in relation to the problem representation, and no specific quantitative manipulation is attempted until a procedure is selected as promising. Therefore, unproductive solution paths would be pursued more often at lower levels of expertise, and solutions are generally less efficient than those carefully planned at higher ability levels. With respect to non-technical problems, direct translation using syntactic processing actually leads to incorrect solution of any problems requiring auxiliary information for accurate comprehension. In general, solvers relying upon the lower levels of solution strategies are limited both in the accuracy and efficiency of their problem solving repertoire.

The progression in problem solving ability sketched in these profiles needs both further empirical exploration and examination for educational implications. Once a developmental schema is firmly delineated, methods for facilitating transition to higher levels of competence may be developed and tested. One preliminary recommendation based on our current understanding of these stages would be to discourage any additional emphasis in the schools on direct translation processes. Because the highest competence levels rely strongly on a Gestalt-like understanding of the entire problem situation, and apprehending relations among problem components is hampered by piecemeal translation, procedures for constructing integrated problem representations need to be encouraged. Secondly, educators might best be sensitive to the solvers' need for well organized knowledge bases, containing concepts related by numerous explicit relations. Schemata comprised of related principles or problem category information appear to facilitate sophisticated problem solving efforts. Finally, it would appear that some means for imparting general procedural knowledge, including strategic approaches and problem solving heuristics, are called for. The highest skilled solvers are most facile at analyzing problem situations, qualitatively elaborating problems with auxiliary information, planning solutions in abstract problem spaces, and comparing features of problems with available schemata. These processes are typically not taught explicitly in educational settings; whether they should be, and if so, how they could be taught, are questions for empirical investigation.

References

Anderson, J.R. Language, memory, and thought. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

Bartlett, F.C. Remembering. Cambridge: Cambridge University Press, 1932.

Bhaskar, R., & Simon, H.A. Problem solving in semantically rich domains: An example from engineering thermodynamics. Cognitive Science, 1977, 1, 193-215.

Bobrow, D.G. Natural language input for a computer problem-solving system. In M. Minsky (Ed.), Semantic information processing. Cambridge, Mass.: MIT Press, 1968.

Brown, J.S., & Burton, R.R. Diagnostic models for procedural bugs in basic mathematical skills. Cambridge, Mass.: Bolt Beranek and Newman Inc., 1977.

Brown, J.S., Collins, A., & Harris, G. Artificial intelligence and learning strategies (BBN Report No. 3642). Cambridge, Mass.: Bolt Beranek and Newman Inc., June 1977.

Buckingham, B.R., & Maclatchy, J. The number abilities of children when they enter grade one. In 29th Yearbook of the National Society for the Study of Education. Bloomington, Ind.: Public School Publishing Co., 1930.

Bundy, A. Analyzing mathematical proofs (or reading between the lines). D.A.I. Research Report No. 2, Department of Artificial Intelligence, University of Edinburgh, May 1975.

Chase, W.G., & Simon, H.A. The mind's eye in chess. In W.G. Chase (Ed.), Visual information processing. New York: Academic Press, 1973.

deGroot, A.D. Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.), Problem solving: Research, method, and theory. New York: Wiley, 1966.

Duncker, K. On problem solving. Psychological Monographs, 1945, 58 (Whole No. 270).

Greeno, J.G. A study of problem solving. In R. Glaser (Ed.), Advances in instructional psychology (Vol. 1). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978a.

Greeno, J.G. Nature of problem-solving abilities. In W.K. Estes (Ed.), Handbook of learning and cognitive processes (Vol. 5). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978b.

Greeno, J.G. Process of understanding in problem solving. In N.J. Castellan, D.B. Pisoni, & G.R. Potts (Eds.), Cognitive theory (Vol. 2). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Greeno, J.G. Indefinite goals in well-structured problems. Psychological Review, 1976, 83, 479-491.

Heller, J.I., & Greeno, J.G. Semantic processing in word problem-solving. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1978.

Hinsley, D.A., Hayes, J.R., & Simon, H.A. From words to equations: Meaning and representation in algebra word problems. In M.A. Just & P.A. Carpenter (Eds.), Cognitive processes in comprehension. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Jurgenson, R.C., Donnelly, A.J., & Dolciani, M.P. *Modern school mathematics: Geometry*. Boston: Houghton Mifflin, 1972.

Kohler, W. *The mentality of apes*. New York: Harcourt Brace, 1927.

Krutetskii, V.A. *The psychology of mathematical abilities in school children*. J. Kilpatrick & I. Wirszup (Eds.), Chicago: University of Chicago Press, 1976.

Larkin, J.H. *Skilled problem solving in physics: A hierarchical planning model*. Unpublished manuscript, University of California at Berkeley, September 1977.

Newell, A., & Simon, H.A. *Human problem solving*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Norman, D.A., & Rumelhart, D.E. *Explorations in cognition*. San Francisco: Freeman, 1975.

Novak, G.S. Computer understanding of physics problems stated in natural language. *American Journal of Computational Linguistics*, 1976, Microfiche 53.

Paige, J.M., & Simon, H.A. Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory*. New York: Wiley, 1966.

Payne, J.N., Zamboni, F.F., & Lankford, F.G., Jr. *Algebra one*. New York: Harcourt Brace & World, 1969.

Reitman, W.R. *Cognition and thought*. New York: Wiley, 1965.

Riley, M.S., & Greeno, J.G. *Importance of semantic structure in the difficulty of arithmetic word problems*. Paper presented at the meeting of the Midwestern Psychological Association, Chicago, May, 1978.

Sacerdoti, E.D. *A structure for plans and behavior*. Menlo Park, Calif.: Stanford Research Institute, Artificial Intelligence Center, Technical Note 109, 1975.

Schank, R.C. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 1972, *3*, 552-631.

Schank, R.C., & Abelson, R.P. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Selz, O. *Die Gesetze der Produktiven and Reproduktiven Geistestatigkeit* (Kurzgefasste Darstellung). Bonn, 1924.

Selz, O. *Uber die Gesetze des Geordneten Denkverlaufs, I*. Stuttgart: Spemann, 1913.

Silver, E.A. *Student perceptions of relatedness among mathematical word problems*. Paper presented at the meeting of the American Educational Research Association, Toronto, March 1978.

Simon, D.P., & Simon, H.A. Individual differences in solving physics problems. In R. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1978.

Wertheimer, M. *Productive thinking*. New York: Harper & Row, 1945. (Enlarged edition, 1959).

Winograd, T. Understanding natural language. *Cognitive Psychology*, 1972, *3*, 1-191.

CHILDREN'S READING PROBLEMS*

Allan Collins
Susan E. Haviland
Center for the Study of Reading
Bolt Beranek & Newman Inc.
Cambridge, Mass. 02138

The question we address in this paper is the issue of why children have problems understanding what they read. At this point, we cannot provide answers that are complete, or even startlingly new, but we can summarize some of the work at the Center for the Study of Reading which we feel is bringing us closer to such answers. Our account of this work will place special emphasis on those aspects that have implications for general educational policy and, more specifically, for testing.

In this paper, we will take up and develop a number of points which we feel are especially relevant to children's problems in reading comprehension. The first is that reading differs from children's early language experiences on a number of different dimensions. We believe that by analyzing these differences we can get some preliminary notions about where children's reading difficulties are likely to occur. The second point is that reading is a process which involves constructing hypotheses based on prior knowledge. Children often develop incorrect hypotheses because they lack some critical piece of information. The third point is that our school system, with its emphasis on decoding skills, engages children in a variety of reading activities that are essentially meaningless. Some children are therefore prone to think that there is no purpose to reading. The fourth point is that once children have learned decoding skills, they are then suddenly faced with functional reading tasks such as reading textbooks or following instructions. An entirely new set of strategic skills is needed for these tasks and no foundation has been laid for them. These four points represent four different areas in which problems in reading comprehension can arise.

Problems Arising from Different Language Experiences

Many of the problems children have in reading depend on the sorts of language experiences they are familiar with before they learn to read. Most children are primarily familiar with conversation, but reading, especially the reading of stories, is quite different from conversation. Some work by Andee Rubin (in press) makes these differences clear. Rubin has isolated a number of dimensions along which language experiences can vary. The dimensions that are particularly relevant to this paper are medium dimensions, dimensions that have to do with how a message is communicated. These dimensions and their values for face to face conversation (c) and reading a story (r) are listed below. For ease of exposition, we call speakers and writers senders and listeners and readers receivers.

---

*The writing of this paper was supported by the National Institute of Education under Contract No. US-NIE-C-400-76-0116. The paper attempts to integrate the thinking of a large number of people associated with the Center for the Study of Reading. In particular there are ideas included from among others Richard C. Anderson, Thomas Anderson, Ann Brown, John Seely Brown, Bertram Bruce, John Frederiksen, Jean Osborne, Rand Spiro, and Andee Rubin.

Modality:
  Spoken (c)
  Written (r)
Interaction:
  Receiver can become sender (e.g. by asking a question) (c)
  Receiver cannot become sender (r)
Specificity of Audience:
  Message designed for particular receiver (c)
  Message designed for generalized receiver (r)
Spatial Commonality:
  Sender and receiver in the same spatial context (c)
  Sender and receiver in different spatial contexts (r)
Temporal Commonality:
  Sender and receiver in the same temporal context (c)
  Sender and receiver in different temporal contexts (r)
Extra-linguistic Communication:
  Gestures and facial expressions possible (c)
  Gestures and facial expressions not possible (r)
Concreteness of Referents:
  Referents visible to receiver (c)
  Referents not visible to receiver (r)
Separability of Characters:
  Statements easily assigned to the character or participant who produced them (c)
  Statements not easily assigned to the character or participant who produced them (r)

These eight dimensions define a multidimensional space in which different sorts of language experiences, such as having a conversation, watching TV, or reading a story, can be represented as points. It is clear that conversation and reading a story are maximally different. In contrast to reading a story, conversations are spoken, the receiver and sender are temporally contiguous and regularly switch roles, and utterances are designed specifically for the participants. Unless the conversation is over a telephone, the participants share the same spatial and temporal context and can also communicate via extra-linguistic means. In most children's conversations, with adults as with other children, the things talked about tend to be concrete and visible (Nelson, 1974). Finally, the contribution of each participant is clearly marked as to its source, and the physical presence of the conversants provides physical anchors for different points of view.

We can illustrate a few specific problems that arise out of these differences. In text, punctuation indicates the higher-level syntactic information that prosodic features, such as intonation and stress indicate in conversation. Thus, a question mark in text corresponds to a rising intonation in speech. A period corresponds to falling intonation, and commas correspond to pauses. However, these correspondences are by no means exact. For one thing, the same punctuation mark can be used for a variety of different purposes, some of them having no analogy in spoken language. For example quotation marks are primarily used to set off quoted material or to indicate a verbatim utterence, but quotation marks can also be used to indicate that the writer is talking about a word or phrase rather than using it, as in the phrase " 'quotation' begins with a q". They can also be used to "hedge" the use of a word or phrase. These uses of quotation marks are peculiar to written language. Since the mapping from punctuation to prosodic features is not exact and since there are higher level syntactic features present in written language that are not present in spoken language, a beginning reader must learn an incredibly complex set of new syntactic skills. It is clear from our work with young children that they have difficulty doing so.

Another problem has to do with the separability of characters. In written stories, the utterances of different characters are set off by quotation marks. Sometimes, in addition there are such phrases as "Joe said ...," but in setting down a long conversation, such phrases become unwieldly and are usually dropped after the first few turns. As result, it is easy to lose track of who said what. Not surprisingly, this often leads to problems in understanding.

Lack of spatial and temporal commonality between the writer and reader also leads to problems, especially with respect to deictic terms such as "here" and "now". In a shared spatio-temporal context, for the most part, what is "here" for the sender is "here" for the receiver and, certainly, what is "now" for the sender is "now" for the receiver. Furthermore, words like "this" and "that" are usually accompanied by gestures indicating exactly what "this" and "that" refer to. In stories, however, deictic words become particularly problematic. "Here" no longer has anything to do with the receiver's spatial location. In addition, unlike other situations which lack spatial commonality, such as telephone conversations or letters, it has little if anything to do with the sender's (author's) spatial location as it rarely if ever refers to the author's location at the time the story was written. Usually it refers to some location within the world created by the story, a location that may not even actually exist. Furthermore, there are differences in meaning depending on how the word is used. When "here" is used by a character, it refers to the character's location. However, when "here" is used as part of the narration it refers to the location in a story currently being talked about or described. Similar sorts of complexity attend the interpretaton of other deictic words such as "now", "this", "that", and even "yesterday", "today", and "tomorrow".

In Rubin's (in press) dimensional analysis there is a maximal separation between conversation, the kind of language experience a child is most familiar with, and reading a story. They have different values on every dimension. Unless children are also familiar with other language experiences that lie along a path between conversations and reading of stories (such as hearing stories or passing notes), they are likely to find reading stories a difficult and unfamiliar experience. Reading stories is difficult not just because of the difficulties inherent in learning how to decode written messages, but also because of the difficulties inherent in simultaneously transferring from one value to another on seven other dimensions as well. Children from cultures in which oral, interactive conversation is especially prevelant, such as Black culture, may lack the necessary familiarity with other sorts of language experience and may find the reading of stories particularly foreign and, perhaps, inexplicable. The problem is, of course, not simply due to strangeness or unfamiliarity. The problem lies in the fact that different values on dimensions require different processing strategies, strategies which most children simply do not have available.

Problems Arising from Lack of Prior Knowledge

Reading involves constructing an interpretation of a set of actions or events in a text based on one's prior knowledge. We can illustrate just how important that prior knowledge is with Schank & Abelson's (1975) example of the knowledge we have about eating at a restaurant. This knowledge is represented in what Schank & Abelson call a script. Most white, middle class Americans have a restaurant script that allows for considerable variation. Consider the differences between eating at a fancy restaurant and eating at McDonald's. When eating at a fancy restaurant you telelphone in advance to make a reservation; you wait for a maitre d' to seat you at a table; you might have a drink while you wait; you are brought a menu at your table; you have several courses which are served by a waiter, etc. None of these events occurs at McDonald's. Therefore reading about them may not make a great deal of sense to children whose restaurant experience is limited to eating at McDonald's. Thus a story about eating at a place like McDonald's will be easier for them to understand than a story about eating at a fancy restaurant.

Recently there have been a large number of psychological studies (Anderson, et al 1978, Bransford & Johnson, 1973) demonstrating that much of what people construct as an interpretation of what they read depends on some critical piece of data. For example, the sentence, "The notes were sour because the seams were split." makes no sense, unless you know about bagpipes. In a similar way children often cannot construct a sensible interpretation, or may construct a wrong interpretation, because they lack a necessary piece of data.

Not only does the interpretation of a text depend on prior experience of such things as restaurants and bagpipes, it also depends on knowledge of the plans and motivations of other people. For example, in stories and fables the characters often talk about one plan, while covertly trying to carry out another (Bruce & Newman, 1978). In one fable about a fox and a rooster, the fox tries to lure the rooster out of a tree by inviting him to breakfast. The rooster accepts on a condition that he can bring along his friend who is in the tree trunk. The fox, thinking he'll have two roosters to eat, invites the friend. Unfortunately for the fox, his friend happens to be a dog rather than another rooster, and the fox is bitten for his trouble. Adults universally interpret this tale as a clever ploy by the rooster to outwit the fox, but usually children do not consider the rooster to be so clever, just lucky (Bruce, 1978). In fact roosters in real life are not so clever, but we as adults have come to expect this type of sophistication of roosters in fables. Here we see how correct understanding of a text can depend on a critical piece of knowledge about a character's motivations.

There are at least two practical implications of this work. One is for children from cultures different from that of the author. To the degree that texts used with children assume familiarity with middle-class situations or motivations, children without middle-class backgrounds will have problems making sense of them. Such children simply won't have the necessary prior information. Any other problems they may have with reading, such as decoding of words will only be compounded. Decoding problems can be compensated for by familiarity with the ideas in the text, since guesses about different words are then more likely to be correct. Therefore, the teaching of reading is made easier by texts that reflect the knowledge and background of the child. Furthermore, testing of reading skills is only unbiased to the degree that the background knowledge of the children being tested is equally appropriate to the texts used.

The other implication is also relevant to the testing of reading comprehension. It is simply that most reading tests are largely measuring background knowledge and not reading skill. It is certainly true that without decoding skills a child cannot understand a text no matter how familiar the ideas may be. However, higher level reading skills, which we will discuss below, are so tightly interwoven with background knowledge that tests must be extremely carefully constructed to separate the two. Thus, reading comprehension tests as currently constructed will inevitably be highly correlated with IQ tests, because both measure background knowledge.

There are, however, a variety of skills specific to reading that one might want to test, beyond word decoding skills. These include the ability to understand the conventions of punctuation and paragraphing, the ability to find specific information in a text, the ability to recognize and recover from wrong hypotheses about the text, and the ability to recognize and use high level text structure. In current tests these reading skills are completely confounded with each other and with background knowledge, but there are several ways in which it is possible to unconfound them. One is to design tests around experiences and motivations that are common to all children taking the test. This least common denominator approach depends on finding such commonalities for its viabiltiy. A second approach would be to use tests that are individually tailored to each child. This could be done by first identifying what the child's interests are or what the child's prior knowledge is. Test items could then be selected to match the child's interest or background. If we want to have truly diagnostic tests this tailored testing approach is the one we will most likely have to take.

There is another way that background knowledge can lead to test bias. Spiro, et al, (1978) have found that children have different strategies for dealing with problems in decoding. One strategy is to try to sound out difficult words in order to identify them. Unfortunately, this technique is slow and can lead to a failure to integrate ideas, because all the effort goes into the processing of individual words. The other strategy is to use the first letter or two to guess at difficult words on the basis of what one thinks they might be from prior knowledge and context. This strategy often culminates in wrong guesses and an inability to recover from incorrect hypotheses about the meaning of a text. Some reading tests make use of cloze procedures in which the child has to fill in missing words in the text. Such a procedure is obviously biased in favor of those students who use the latter strategy, even though the former strategy may be better in the long run. A fruitful direction for test design may be to attempt to distinguish these two classes of children and teach them to alternate strategies when appropriate.

Problems Arising from the School System

Several aspects of the way schools in America teach reading may cause problems for children. These problems derive from too much stress on the decoding of words and in particular from the currently popular componential skills approach to reading.

In the early grades of school, children are faced with a variety of activities that are designed to teach componential skills in reading, for example, tasks such as circling pictures whose words begin with b. For the most part these activities are tedious, and they are such fractionated language experiences that they seem pointless to the children engaged in them. They are one of the ways we inadvertantly teach children that reading is not inherently pleasurable, but something done to please the teacher. Students who don't care about pleasing the teacher will not care about doing these tasks.

In contrast to this fractionated approach, it is possible to engage children in language activities that are meaningful and also stress various component skills. Many such activities are suggested by the Rubin taxonomy described above. If a child is having difficulty with one or another dimension of reading, he or she can be given an activity that is like reading on that dimension, but more like conversation on other dimensions. For example, having stories read aloud to a child teaches notions about text structure while easing the decoding problem. Having children pass notes back and forth teaches decoding and writing skills while maintaining most of the aspects of conversation. There are no problems with deixis or with separability of characters in such an activity. Computer games that involve reading, such as Adventure, teach decoding skills and the following of instructions in an interactive and personalized way. Having children play these games provides a painless way of exercising such skills. The above are just a few examples of reading activities that are highly motivating and that stress componential skills while still maintaining the communicative function of language.

Perhaps the most destructive example of giving children tedious activities is what happens when children have trouble with decoding. Usually such children are given remedial texts that stress certain phonic patterns. For example, one such text is full of words like "rag", "bag", ragbag", "tag", etc. and another emphasizes "oa" words like "boat", "road", etc. The striking thing about these texts is that their meaning is incredibly difficult to figure out, if indeed there is any meaning at all. In effect they teach poor decoders that reading is a meaningless activity; reading is saying the correct words, rather than understanding them. In their immediate attempt to teach decoding skills the designers of these texts seem to have forgotten the ultimate goal of reading.

With the new techniques for analyzing the structures of stories developed by such investigators as Rumelhart (1975) and Bruce & Newman (1978), it is now becoming possible to evaluate the meaningfulness of texts. We can now rate stories on how interesting their plot structure is, how well they incorporate conflict or surprise, and how coherent they are. If such ratings were made of some of the beginning texts now used in schools, these texts would be found sadly lacking. Until now we could only make readability evaluations in terms of vocabulary and sentence length. Because of this, texts were produced that were readable by these standards, and unreadable by other standards. What was being taught had been reduced to what could be measured.

Problems Arising from the Transition to Functional Reading

Comprehension difficulties often arise in the third and fourth grade because children move from reading texts designed to teach basic reading skills to reading texts designed to convey information (e.g. social study texts or homework instructions). The beginning reading texts use vocabulary and situations that are reasonably familiar to most children, but the informative texts involve new ideas and new vocabulary. As the task of reading changes from acquiring decoding skills to acquiring new information, a new set of strategies becomes relevant. Among them are how to deal with failures to understand different words and phrases, how to revise misinterpretations, and how to select important points for whatever purpose is at hand (such as doing a task if given instructions, or remembering later if given expository material). These strategies represent a set of skills that children have not needed before and for which they have received no preparation in the earlier grades.

To some extent the same problems arise in conversation, but the appropriate strategies for dealing with such problems are quite different from those needed in reading. When you don't understand something in a conversation, you can look puzzled or say "what?" Such techniques do not work in reading. Hence, children do not have strategies for dealing with such problems available and for the most part, they are never explicitly taught such strategies. In such situations, the brighter children develop their own strategies and the less bright children lose interest.

We can illustrate the new problems children face with three examples:

1) What to do when you don't understand.
2) How to read to remember later.
3) How to recover from wrong hypotheses.

Our preliminary work (Collins, Brown, Morgan & Brewer 1977; Collins, Brown & Larkin, in press) indicates that skilled readers acquire a variety of tacit strategies for what to do when they don't understand a text. In such cases they appear to carry along a set of questions as they read. If the structure of the text suggests these questions will be answered later, skilled readers will continue to read. If, however, too many questions collect, they will often jump back to the sentences that led to the questions. These they reread in order to form hypotheses which allow them to cut down the number of open questions. Learning when to keep reading, and when and where to jump back and reread are skills that are crucial for reading difficult material.

Reading for memory is another skill that only arises in later grades when teachers begin to expect children to acquire information on their own. It involves several subskills that students need to acquire, such as picking out main points, saving them in some form by underlining or rehearsing, and skimming the same text later. Brown & Smiley (1977) have extensive data that show that children do not know how to pick out the main points in a text. Because they haven't learned this skill they don't know what to focus their efforts on, and hence their studying is not as fruitful as that of adults.

We have discovered through tutoring children in reading that they tend to hold onto wrong hypotheses even when they encounter contrary evidence. This should be expected from the psychological literature (Bruner & Potter, 1964), but giving up wrong hypotheses is an important skill that good readers must acquire. The appropriate strategy is to go back and reread looking for a new interpretation, rather than to cling steadfastly to hypotheses which no longer make sense.

These three meta-comprehension skills are examples of reading skills that are necessary for survival in school but that are not taught in school. These are all skills that are specific to reading, so they do not arise out of earlier language experiences. It is possible to develop tests to diagnose whether children have acquired these skills, but they must be carefully designed (see section on Prior Knowledge). Where children lack any of these high-level strategic skills we should attempt to teach them directly.

The Relation Between Teaching and Testing

In an ideal world, the function of testing would be to identify the problems a student is having. Such testing would then determine what is taught to the student. This is a highly individualized notion of testing and teaching, one in which the testing provides feedback to guide the teaching.

There has been some movement in the schools away from tests that measure overall reading skills towards tests that diagnose specific problems. This is clearly a move in the right direction, but these tests can be made even more diagnostic by separating out background knowledge and by pinpointing underlying skills. In particular, if we can develop well-articulated theories of how the reading process can go wrong, as Brown & Burton (1978) have done for arithmetic, then we can develop significantly better reading tests. This is one of the Reading Center's major goals.

In conjunction with such tests we still have to develop new materials that can be used to remedy problems that the new tests diagnose. These materials should be centered around interesting and meaningful language activities. The minicomputer revolution will soon allow us to make many of these activities computer-based.

In summary, we believe it is possible to diagnose specific problems arising from a number of different sources that children have in reading. Such diagnostic tests must then be tied to educational activities aimed at remedying the specific reading problem a child has.

References

Anderson, R. & Pichert, J. Recall of previously unrecallable information following a shift in perspective. Journal of Verbal Learning and Verbal Behavior, in press.

Anderson, R. Schema-directed processes in language comprehension. In A. Lesgold, J. Pelligreno, S. Fokkema, and R. Glaser (Eds), Cognitive psychology and instruction. New York: Plenum, in press.

Bransford, J. & Johnson, M. Contextual prerequisites for understanding: Some investigations of comprehension and recall. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 717-726.

Brown, A. Knowing when, where, and how to remember: A problem of metacognition. Center for the Study of Reading, University of Illinois at Urbana-Champaign, Technical Report No. 47, June 1977.

Brown, A., Smiley, S., & Lawton, S. The effects of experience on the selection of suitable retrieval cues for studying from prose passages. Center for the Study of Reading, University of Illinois at Urbana-Champaign, Technical Report No. 53, July 1977.

Brown, A. & Campione, J. Memory strategies in learning: Training children to study strategically. In H. Pick, H. Leibowitz, J. Singer, A. Steinschneider, and H. Stevenson (Eds.), Applications of basic research in psychology. New York: Plenum Press, 1978, in press.

Brown, A. & Smiley, S. The development of strategies for studying prose passages. Center for the Study of Reading Technical Report No. 66, October 1977.

Brown, J.S. & Burton, R.R. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 1978, 2, 155-192.

Bruce, B. What makes a good story? Language Arts, 1978, 55, 460-466.

Bruce, B. & Newman, D. Interacting plans. Cognitive Science, 1978, 2, 193-235.

Bruner, J. & Potter, M. Interference in visual recognition. Science, 1964, 144, 424-425.

Collins, A., Brown, A. Morgan, J., & Brewer, W. The Analysis of Reading Tasks and Texts. Center for the Study of Reading, University of Illinois at Urbana-Champaign, Technical Report No. 43, April 1977.

Collins, A., Brown, J. & Larkin, K. Inference in text understanding. In R. Spiro, B. Bruce, & W. Brewer (Eds.), Theoretical issues in reading comprehension, Hillsdale, N.J.: Lawrence Erlbaum Associates, in press.

Nelson, K. Concept, word, and sentence: Interrelations in acquisition and development. Psychological Review, 1974, 81, 267-285.

Rubin, A. Comprehension processes in oral and written language. In R. Spiro, B. Bruce, & W. Brewer (Eds.), Theoretical issues in reading comprehension, Hillsdale, N.J.: Lawrence Erlbaum Associates, in press.

Rumelhart, D.E. Notes on a schema for stories. In D. Bobrow & A. Collins (Eds.), Representation and understanding: Studies in cognitive science. New York, Academic Press, 1975.

Schank, R. & Abelson, R. Scripts, plans, and knowledge. In Proceedings of the Fouth International Joint Conference on Artificial Intelligence. Tbilisi, Georgia: USSR, 1975, 151-157.

Spiro, R. & Smith, D. Distinguishing sub-types of poor comprehenders: Overreliance on conceptual vs. data-driven processes. Center for the Study of Reading, University of Illinois at Urbana-Champaign, Technical Report No. 61, October 1977.

DEVELOPMENT IN WRITING*

Carl Bereiter
Ontario Institute for Studies in Education

Although there is a substantial body of data on the development of writing skills, it has not seemed to have much implication for instruction. Reviews of writing research from an educational perspective have given scant attention to it (Blount, 1973; Braddock, Lloyd-Jones & Schoer, 1963; Lyman, 1929; West, 1967). Generally speaking, developmental research has educational significance only when there is a conceptual apparatus linking it with questions of practical significance.

Almost all of the data on writing development consist of frequency counts--words per communication unit, incidence of different kinds of dependent clauses, frequency of different types of writing at different ages, and so on. The conceptual frameworks used for interpreting these data have come largely from linguistics (e.g., Hunt, 1965; Loban, 1976; O'Donnell, Griffin, & Norris, 1967). However informative these analyses might be to the student of language development, they are disappointing from an educational point of view. The variables they look at seem unrelated to commonly held purposes of writing instruction (Nystrand, 1977).

The purpose of this paper is to synthesize findings on the growth of writing skills within what may be called an 'applied cognitive-developmental' framework. Key issues within an applied cognitive-developmental framework are the cognitive strategies children use and how these are adapted to their limited information processing capacities (Case, 1975, 1978; Klahr & Wallace, 1976; Scardamalia, in press). Although this paper will not deal with instructional implications, it will become evident that the issues considered within an applied cognitive-developmental framework are relevant to such concerns of writing instruction as fluency, coherence, correctness, sense of audience, style, and thought content.

What Is Development in Writing?

Students' writing will undoubtedly reflect their overall language development (Loban, 1976; O'Donnell et al., 1967) and also their level of cognitive development (Collis & Biggs, undated; Scardamalia, in press). It is to be expected that it will also at times reflect their level of moral development, social cognition, etc. Given the small role that writing plays in most children's lives, it is therefore reasonable to suppose that there is no such thing as writing development as such--that it is merely the resultant of other, more basic kinds of development. This view is implicit in the speech-primacy position of researchers like Loban (1963, 1966, 1976), who treat children's writing simply as another source of data on their language development.

---

*To be published in Gregg, Lee W. and Steinberg, Erwin R., eds. Cognitive Processes in Writing. Hillsdale, N.J.: Erlbaum, 1979

While it may be reasonable to treat writing development as a reflection of other kinds of development, it is not very useful to do so. An educationally relevant account of writing development would have to give prominence to whatever is distinctive about writing and potentially susceptible to direct influence. The following have been recognized as distinctive characteristics of writing.

1. Written English may be recognized as a subsystem of English, along with spoken English, and distinguishable from the latter in a number of ways. It is usually more compact, contains more elaborately specified subjects, and shows less local variation than spoken English, and generally shows a different distribution of linguistic devices and usages (Allen, 1972; Gleason, 1965; Long, 1961).

2. Written English and spoken English are predominately, but not exclusively, tied to different modalities. One may speak written English, as in dictation, or write spoken English as in composing dialogue. Unpublished data collected by P.J. Gamlin show that children as young as age eight can differentiate sharply between writing spoken English and writing written English, although there is a tendency for them to drift into writing spoken English. Use of the two subsystems is evidently quite situation-sensitive. Moscovici and Plon (1966) found that when people are seated back-to-back, their conversational speech takes on the characteristics of written language, whereas it has the characteristics of spoken language when they are seated face-to-face, even when there is a screen between them.

3. Written English involves a number of conventions not pertinent to spoken English-spelling, punctuation, paragraph indentation, etc. What is especially noteworthy about these conventions, as compared to the conventions of spoken language, is that they are learned almost exclusively in school (Allen, 1972).

4. Writing may require and thus foster a different kind of thought from that involved in speaking (Goody & Watt, 1963; Havelock, 1973; Olson, 1977). Because writers get little feedback and because they must address unknown readers in unknown contexts who have unknown states of knowledge, writers are forced to develop what Olson (1977) calls "essayist technique"--a form of discourse characterized by explicit reference and by propositions linked together by logical entailment rather than by reference to experience shared between writer and reader.

5. Finally, because written utterances can be more deliberately shaped and reshaped than spoken ones, writing lends itself to the development of craftsmanlike skills not normally found in speaking. For the same reason writing also lends itself to complex productions that would be very difficult to achieve in the oral mode--for instance, novels, sonnets, and scholarly works.

The preceding observations suggest that a useful account of development in writing should deal with (1) the gradual differentiation of written from spoken language, as distinct subsystems, (2) development of ability to switch appropriately between the systems, (3) mastery of the conventions peculiar to written language, (4) development of explicit, objective, context-free propositional language, and (5) achievement of literary style and proficiency in various genres of written composition.

Studies comparing spoken and written language development have indeed shown a gradual differentiation of the two. In the early school years the mean T-unit length[1] of spoken utterances is greater than that of written ones and somewhere in the middle school years this difference reverses, in keeping with the tendency of mature written language to be more elaborated than spoken language (Loban, 1976; O'Donnell et al., 1967).

Data more directly indicative of mastery of writing skills have not shown such regular growth, however. Global ratings of quality of written compositions in the National Assessment of Educational Progress (1975) show that while substantial gains in quality occur between the ages of 9 and 13, little or no gain occurs between the ages of 13 and 17. Hill (1972), on the other hand, found no natural growth or development in the use of specifically literary devices of genre or form; all devices could be found in children's writing, but with a low incidence at all ages.

While Loban (1976), in his 12-year longitudinal study of written and oral language development, found steady trends in gross indicators of quantity and complexity of verbal output, trends in the use of specific language devices tended to be more wobbly. For instance, while the use of dependent clauses in oral language increased steadily from year to year and always distinguished high ability from low ability students, the same was not true in written language. For students of high and average ability, use of dependent clauses leveled off after grade 8, and low ability students overtook them by grade 10. Loban's interpretation was that by that time the more able students were finding alternative means of embedding. In general it seems that frequency-count indicators of growth become equivocal when applied to specific usages, because linguistic devices are often substitutable one for another, so that increase in the use of one may result in decrease in the use of another. Quality indicators may also show unstable trends because of regressions due to the introduction of new skills into the writer's repertoire. Diebel and Sears put it well decades ago when they noted that some errors in writing increase with age because "as the child's mind grows he will use more complicated forms of expression" (Diebel & Sears, 1917, p. 175).

These ambiguities in frequency-count findings suggest that lying behind the quantitative trends in writing development may be more interesting qualitative changes, changes in the way children go about writing. Research aimed at discovering such qualitative changes has been rare, however.

Loban (1976) stated as one of the purposes of his longitudinal study to investigate the question, "Can definite stages of language development be identified?" (p.2). In the end he was able to provide qualitatively distinct descriptions of the language of children at different ages, but what they nevertheless generally add up to is the gradual acquisition of mature characteristics. As Loban (1976, pp. 84-85) observed, his data "very often show a steady nondramatic chronological development. This would indicate that linguistic 'stages' are not more discrete, no more sudden, than the stages of physical growth reported by Gesell and Ilg." In none of his 'stage' descriptions does Loban note anything special about writing development. He does, however, observe that there is a plateau in oral language development runnning from grade 7 to grade 9 or 10 and that this plateau comes a year later in written language.

The most ambitious attempt to identify qualitative changes in writing abilities is the Schools Council study of English students of ages 11 to 18 (Britton, Burgess, Martin, McLeod & Rosen, 1975). This research was based on samples of writing done in all curriculum areas. Compositions were classified according to a complex model of written language functions and writer-reader relationships. Without explicating the model, which is impossible in this review, the findings cannot be adequately conveyed. One major finding, however, stands out above all the complexities. It is the over-riding influence of school demands on student writing. Although there was some evidence of a developmental trend toward writing for a larger audience, it was overshadowed by a trend (presumably school-induced) toward writing to the teacher as examiner. The authors hypothesized a branching developmental trend from relatively undisciplined "expressive" writing to "transactional" and "poetic" writing. Data generally supported the hypothesis, except that there was a drastic decline in the frequency of "poetic" writing in the oldest age-group. This shift was attributed, not to natural tendencies, but to the influence of the examination system. Thus, the disappointing message of the Schools Council research is that as soon as we begin to look beyond syntax, vocabulary, and the like and try to investigate functional aspects of student writing, we begin to find out more about the school system than we find out about children.

A research approach that shows promise of getting at qualitative changes in writing competence has been reported by Stahl (1974, 1977), who has devised an instrument for describing structural characteristics of compositions on a particular topic--"My Home." Applying this instrument to compositions written by Israeli children in grades 2, 5, and 8, Stahl (1977) obtained results that could be interpreted as showing a stage shift between grades 2 and 8 (although Stahl interpreted them as showing gradual improvement). The modal compositions in grades 2 and 8 were alike in showing no evidence of planning or structural revision and no indicated order of presentation. However, the model grade 8 composition followed a general principle for selecting content, presented the content in a coherent ("guided tour") arrangement, and allocated space in proportion to the importance of elements, whereas the modal grade 2 composition showed no evidence of attention to selection, arrangement, or balance. The typical grade 8 composition ended at an appropriate point, whereas the typical grade 2 composition just broke off. It is the grade 5 compositions, however, which provide support for a stage shift interpretation of these differences. Instead of falling into intermediate categories between those occupied by the grade 2 and grade 8 compositions, they tended to be bimodally distributed, some showing the grade 8 characteristics and others showing the grade 2 characteristics.

What is encouraging about Stahl's approach is that it begins to tell us something about the writer instead of about the writer's product. It begins to suggest a different and more elaborate system of skills operating in grade 8 students from that operating in grade 2 students. Nonetheless Stahl's, along with all the other developmental research cited, is what Nystrand (Note 1) calls 'text-based' as opposed to 'writer-based.' It looks for what is achieved in the written product rather than using the written product to infer what is inside the writer--that is, what constitutes the writer's competence or executive system.

However, although "Look at the process, not the product" is a fine precept, writing researchers of the past should not be criticized for their failure to follow it. Following this precept, and thus advancing toward an understanding of how the process of writing develops in children, requires a psychological model of the writing process; and work on developing cognitive process models of writing has only recently begun.

Writing as a Cognitive Process

Although no complete cognitive processing model of writing has yet been advanced, several beginnings have been made. Hayes and Flower (1978) have been extracting, from protocols of writers 'thinking aloud,' the basic moves or strategic elements that writers use in written composition. Nold (in press), has attempted to sort out different levels of processing that go on in revising, levels which clearly have their counterparts in composition. Scardamalia (in press) has looked specifically at how immature writers adapt writing tasks to fit within their limited information processing capacities. A complete processing model would have to deal with all three of these aspects--with the cognitive moves that make up writing and their organization; with levels of processing, from the highly conscious and intentional to the unconscious and automatic; and with how processing capacity is deployed to these various functions in such a way as to enable writing to go on.

Since an understanding of the development of writing abilities depends on such a model, it is worth considering what a complete model of the writing process might be like, even if it can be sketched only roughly. A sketch is possible even at this early stage of cognitive research into writing because a model of written composition will necessarily have much in common with models of speech production (e.g., Fodor, Bever, & Garrett, 1974) and of language comprehension (e.g., Adams & Collins, 1977).

There will be a high-level executive scheme directing the whole writing operation in keeping with certain purposes and constraints. At the next lower level will be what in writing we may wish to call 'genre' schemes. A genre scheme consists, essentially, of the knowledge available for directing a certain kind of writing. The person who has never before written a mystery story is likely nevertheless to have available a mystery story scheme that will direct the novice writer to think of a crime, a crime-solver, a way of bringing the solver into contact with the crime, a puzzling element, etc., that will call for keeping the identity of the criminal secret till the end, and that may call for certain stereotyped events, characters, and expressions. The experienced mystery writer will have a more elaborate genre scheme that will set more sophisticated requirements and direct more complex searches and constructions.

In order to consider the functions of a genre scheme, let us take a more everyday example, that of a 'letter of recommendation' scheme. The teacher or supervisor experienced in writing such letters will have a scheme that includes the following:

1. A limited set of fairly specific intentions: e.g., to present the candidate in the best possible light, to do the candidate no harm but to avoid explicit commitment, etc.

2. A set of strategies or 'game plans' appropriate to carrying out these intentions. A game plan will in turn include the following.

3. Categories of content needed to support the plan. These may be in the nature of slots to be filled: "The candidate has shown initiative by __________ and __________."

4. Search procedures for discovering the needed content. These may be overt procedures such as consulting certain records or calling on informants, or they may be internal memory-search strategies.

5. Tuning instructions for language output. For instance, a typical game plan might call for expressing the recommendation in language that is standard written English, fairly formal, dense, authoritative, and vague. But another strategy might call for quite different language.

The beginner, who has never before written a letter of recommendation and who may not even have read one, will lack a specific scheme for this genre and will have to use some other available scheme--such as a general business letter scheme or a scholarly writing scheme. These schemes will also contain intentions, game plans, categories or content, search procedures, and tuning instructions, but they will be less finely adapted to the task of writing a letter of recommendation. That means that a great deal more feedback and revision of higher-level decisions in the light of lower-level outputs will be required. If the beginner is a generally skillful writer, the result may well be an adequate letter of recommendation, but it may take the beginner ten times as long to achieve the same result as it takes the writer who has a well developed scheme for that genre.

Below the genre scheme will be a content processor, which draws semantic material from memory and organizes it according to instructions from the genre scheme. Its output is what we may call a gist--a unit of content not yet formed into overt language. The gist then goes to a language processor which puts it into explicit language, according to tuning instructions from the genre scheme. The content and language processors, we may assume, are of an all-purpose sort and are not specialized for genres nor even to writing as opposed to speaking. What they do that is unique to writing depends on instructions from writing genre schemes.

This simple step-at-a-time process is sure to be enormously complicated by continual comparisons between instructions and outputs, with resulting changes in processing or in higher-level decisions or in both. Without going into detail, we may simply note the extreme case, which is likely to be familiar to any experienced writer--the case where difficulty in finding the right word starts a chain of adjustments at successively higher levels until finally the intention of the whole composition is altered.

By the time children have acquired the basic mechanics of writing, they have probably available the whole hierarchy of processing levels that adults have, even though the schemes or mechanisms at each level may be undeveloped and little differentiated. It is clear that young children can formulate intentions (J. McV. Hunt, 1963). They can broadly differentiate their writing according to genres such as narrative, dialogue, exposition, and verse (Hill, 1972). And they can tune their language output to suit different audiences (Shatz & Gelman, 1973).

One interpretation of writing development might therefore be that it consists of the gradual elaboration and refinement of relevant schemes at different processing levels. This is the view that seems to be implicit in most school approaches to writing instruction, where pupils at all grade levels do pretty much the same thing--write stories, essays, etc.--and the teacher tries to help students along with all aspects of writing at once. Such a view is undoubtedly correct in that learning inevitably takes place at all levels at once. For an analysis of writing development to stop there, however, is for it to overlook all that is most difficult about learning to write.

Information Processing Load in Writing

The preceding sketch of a process model of writing only weakly suggests the number of things that must be dealt with more-or-less simultaneously in writing. As Scardamalia (in press) observes,

> Even a casual analysis makes it clear that the number of things that must be dealt with simultaneously in writing is stupendous: handwriting, spelling, punctuation, word choice, syntax, textual connections, purpose, organization, clarity, rhythm, euphony, the possible reactions of various possible readers, and so on. To pay conscious attention to all of these would overload the information processing capacity of the most towering intellects.

Although some of these matters could be dealt with sequentially instead of simultaneously, it appears that most writers do not do so. Emig (1971), questioning mature and practicing writers, found few who planned before writing, developed text from outline, etc. Stallard (1974), examining the writing behavior of "good" as well as randomly selected student writers, again found that both kinds tended simply to plunge into the task and try to do everything at once. Our own informal observations of careful writers suggest that they do most of their revising as they go along, instead of treating revision as a separate segment of the writing process.

In order for writers to carry on such a variety of processes simultaneously, either or more likely both of the following conditions must obtain: (1) automatization of many parts of the writing process so that they can be carried on with infrequent or slight conscious attention (Posner & Keele, 1973), and (2) highly skilled time-sharing, so that attention can range over a number of on-going tasks without serious lapses or interference. Such performance would also seem to require attainment of adult levels of working memory capacity--if indeed this capacity grows with age (Pascual-Leone, 1970; Simon, 1972, 1974).

The only known attempt at a quantitative analysis of information processing load in writing is that of Scardamalia (in press), and her analysis looked at only one aspect or processing load--the coordination of content schemes in expository writing. Even so, her findings are of considerable developmental interest. Among high-achieving children of ages 10 to 14, the treatment of content appears to be at a very low level of complexity. Children who are presumably capable of handling from three to five content schemes simultaneously tend in writing to manage only one to three. This would suggest that they are so hampered by low-order problems of getting language onto paper that they have little attentional capacity left over for higher-order concerns with content.

A useful theory for interpreting the information processing load aspects of writing is Pascual-Leone's theory of constructive operators (Pascual-Leone, 1970; Pascual-Leone & Goodman, 1976; Pascual-Leone & Smith, 1969). Of special importance as far as the analysis of writing abilities is concerned are two features of the theory: (1) it provides analytical methods that permit quantification of information load in the same terms over different kinds of task content (Bereiter & Scardamalia, in press) and (2) it brings under the same conceptual and analytical umbrella different ways that cognitive elements may secure a place in working memory. This second feature is important in explaining what would otherwise be anomalies in children's writing performance. Menig-Peterson & McCabe (1977) have observed that, while young children tend to produce incomplete and unfocussed narratives under ordinary task instructions, they frequently produce complete and well-integrated narratives when stimulated to tell about an event of strong emotional import.

According to Pascual-Leone's theory, the principal limitation on the complexity of cognitive performance is the number of mental schemes that a person can keep simultaneously activated. A certain limited number can be kept active through mental effort--a number that increases with age, ranging from 2 to 7 across the span of school years. But additional schemes may become activated because of emotional loading or because they are triggered by salient stimuli in the immediate environment. Thus it is that under facilitating emotional and stimulus conditions children can sometimes perform in ways characteristic of older children or adults (Odom, in press; Odom & Cunningham, 1975; Scardamalia, 1975). There is a tendency, as manifested in the Menig-Peterson and McCabe (1977) paper, and clearly traceable to the influence of Labov (1970), to treat the very best performance that can be elicited from children as defining the norm and to attribute everything else to experimental blundering or bias. When applied to a poorly understood phenomenon like writing behavior, such a view could seriously distort any normal developmental trends that might exist.

On the basis of the preceding considerations, we may state the following as a fundamental constraint on both the form and pace of writing development: Mature writing involves a large number of skills at different processing levels. Adequate mature functioning can be possible only when many of the skills are highly automated and when they are well enough coordinated to permit efficient time-sharing. Neither of these conditions is met in the young writer, and so the young writer, in order to function at all, must employ a structurally simpler system that does not require so much simultaneous and coordinated functioning. Since low-order schemes--those involved in getting words on to paper--must take priority in order for writing to occur, it follows that the system employed by the young writer must be one in which low-order schemes predominate and higher-order schemes play a lesser part (Shuy, 1977). Movement toward a more complex system, in which higher-order schemes predominate, would have to wait until the lower-order schemes are sufficiently automatized.

At present there is not sufficient evidence by which to choose between a "gradualist" and a "structuralist" conception of writing development. A gradualist conception would hold simply that as low-order skills become automatized, higher-order skills come into use. A structuralist conception would hold that the writing process, however it is carried out, has organization and that the incorporation of a new skill requires reorganization of the process. Accordingly, writing development will not be characterized simply by gradual elaboration and refinement of schemes at different processing levels but by more discrete stages of organization.

Possible Stages in Writing Development

While a structuralist conception of writing development has a certain plausibility, it remains to construct a reasonable structural model--that is, a model that shows what successive structurally distinct stages of writing development might actually be like. A provisional model of this kind is advanced in this section. The word "stage" inevitably evokes the idea of Piagetian stages and all their associated conceptual baggage, but the word is used here with a more limited connotation. A stage is simply a form of organization that is preceded or followed by other forms. To suggest that there may be distinct stages in the development of writing abilities is not to suggest that those stages are universal or that they have a necessary order, much less that they are yoked to the Piagetian stages of cognitive development. The stages to be described here have a seemingly natural order to them, but it is quite conceivable that with a different sort of educational experience children might go through different stages with a different kind of order.

Six different systems of knowledge or skill may be identified in mature writing: fluency in producing written language, fluency in generating ideas, mastery of writing conventions, social cognition (appearing as ability to take account of the reader), literary appreciation and discrimination, and reflective thought. Each of these skill systems can develop, to a greater or lesser extent, independently of the others, and, of course, simultaneous development in all of them is to be expected across the span of school years. This does not mean, however, that all of these skill systems are involved in writing performance across the school years. Rather, it seems that writing development might follow a course similar in form to that which Schaeffer (1975) has described for development of number concepts.

Schaeffer (1975) calls his model of development "hierarchic skill integration." Because of children's information processing capacity limitations, they cannot integrate all skills at once, as they would have to do in order to achieve something formally equivalent to mature performance. Instead, they integrate what skills they can in order to achieve a functional skill of a higher order. As this higher-order skill becomes automatized, they integrate another skill with it to achieve yet a higher-order skill, and so on. Applying this hierarchic concept to the six skill systems mentioned above, we get the picture of hierarchic skill integration shown in Figure 1 and elaborated below.

1. Associative writing. The simplest system capable of producing intelligible writing is that which combines the two skill systems of fluency in written language and ideational fluency. The resulting skill of associative writing consists essentially of writing down whatever comes to mind, in the order in which it comes to mind. It is what Bobrow and Norman (1975) call a data-driven, as contrasted with a concept-driven process. An empirical demonstration of such a process in writing comes from a study recently carried out by Brooks Masterton, Marlene Scardamalia, and the writer. Subjects were shown a picture by having masking pieces removed one at a time. Each exposure added significant new information. In describing the picture, the younger the subjects, the more likely they were to mention first what was newest to them as opposed to what would be most informative to a reader.

Much unskilled writing by students of all ages appears to be of this kind. One of the characteristics of "writer-based prose" noted by Flower (1977) in college students is that ideas are presented in the order they were acquired by the writer rather than in an order suited to the reader. The less talented grade ten students whom we have interviewed even describe their writing as a process of putting down thoughts as they come to mind. Indeed, what they say about writing is much the same as what Keeney (1975) found with intermediate-grade children and Graves (1973) found with children at the very beginning of learning to write: Their principal obstacle is thinking of what to write. Given an 'interesting' topic--that is, one that elicits an immediate flow of ideas--they can write easily and are pleased with what they produce. Self-chosen topics are, as one might expect, more likely to produce fluent ideation than are teacher-chosen ones. If the topic is 'dull,' then writing is difficult and the result, as judged by the students themselves, is a 'dull' composition. This would follow from the data-driven character of the process. The associative writer lacks resources for creating an interesting composition. The same

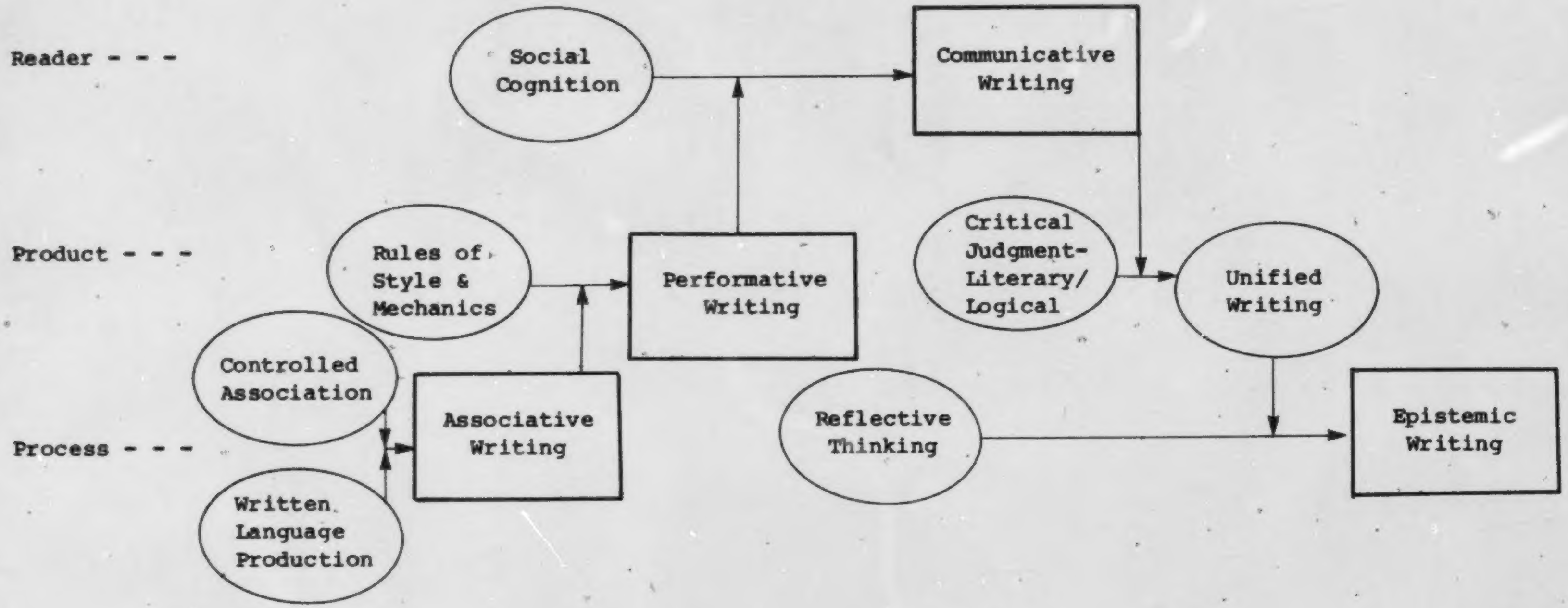

Reader - - -
Product - - -
Process - - -
Social Cognition
Communicative Writing
Rules of Style & Mechanics
Performative Writing
Critical Judgment- Literary/ Logical
Unified Writing
Controlled Association
Associative Writing
Written Language Production
Reflective Thinking
Epistemic Writing

Figure 1. A model of skill systems integration in writing development.

is true as regards coherence and completeness. Associative writers say they stop writing when they run out of ideas. If the topic is emotionally arousing enough that they need devote no mental effort to keeping the topic in mind, then their productions will tend to be unified and to appear fully developed and coherent. With less arousing topics, however, they are liable to lose hold and go off on a chain of associations irrelevant to the topic rather than produce associations that are all related to the topic (Scardamalia, in press).

Britton et al (1975) recognize a similar early stage of writing that is relatively unplanned, uninhibited, unconcerned about audience reaction, and that is on the whole closer to transcribed speech than it is to literate writing. Their term for it is "expressive" writing. While expressive may appropriately describe the spontaneous writing of children, it does not fit their more workaday efforts, which however, display the same characteristics. When 10-year-olds were set the task of writing instructions for a game they had just learned, for instance (Scardamalia, Bereiter & McDonald, in press), many of them exhibited all the characteristics mentioned above, but their writing could scarcely be called expressive.

The similarity of associative writing to transcribed speech may have an objective basis. Simon (1973) observed that children in the first years of schooling typically verbalize or subvocalize what they write, suggesting that they are in fact dictating to themselves. By the third year of school this behavior had largely disappeared from the girls observed by Simon, and by the fourth year from the boys. This would suggest that by middle-childhood the two skill systems of generating written language and controlled semantic association have become integrated into associative writing. Recent data that we have been collecting suggest, however, that this may be true only of narrative writing and that attainment of skillful associative writing comes much later for expository writing.

2. Performative writing. Most of traditional school writing instruction may be interpreted as an effort to shape students' associative writing into conformity with conventions of style and mechanics. This has proved to be a difficult instructional task. Although children may learn to spell and punctuate correctly in workbook and test exercises, thus demonstrating learning of these conventions, it remains a major step for them to incorporate this learning into their normal writing. The same is true for higher-level conventions such as use of varied sentence patterns and avoidance of ambiguous reference. From the student's standpoint, concern about style and correctness disrupts the flow of associative writing. As one of our subjects put it (previously quoted by Scardamalia (in press)), "I have all my thoughts in my mind but when I come to a word I can't spell it throws me off my writting (SIC)."

What is here termed performative writing consists of the integration of associative writing with knowledge of stylistic conventions. It is reasonable to speak of this as a stage because it is a functional system actually achieved by many people--people who can, seemingly without conscious attention, produce writing that consists of well-formed standard English sentences, correctly spelled and punctuated. Performative writing is not, however, a stage in the sense of representing a single massive reorganization of behavior. Rather, it is clear that progress toward performative writing consists of a vast number of skill integrations; each requiring attention to a particular matter until it becomes an automatic part of writing behavior--the spelling of particular words, the punctuation of particular constructions, the use or avoidance of particular expressions, etc. (Shaugnessy, 1977). Thus it is that growth in the performative aspects of writing shows an overall continuity. But each particular skill integration imposes the same kind of cognitive demand on the learner--to attend simultaneously to the content of writing and to some feature of the emerging written product. Children's difficulties with revision--which is usually revision to bring the composition into closer conformity with conventions--undoubtedly also reflect this attentional capacity problem (Nold, in press).

3. Communicative writing. When performative writing is integrated with social cognition, the result is communicative writing--that is, writing which is calculated to have a desired effect on an audience. Britton et al (1975) have used the term transactional to designate this kind of writing, contrasting it with poetic. In making this distinction, however, they limit the former kind of writing to what is more commonly known as expository writing. This categorization may be valid rhetorically: clearly, different performative skills are involved in convincing or informing a reader from those that are involved in amusing or wringing tears from a reader. But from the standpoint of skill systems integration, they are all part of the same family.

Social cognition has a complex developmental history of its own (Looft, 1972; Shantz, 1975). Even toddlers appear capable of adapting messages to salient characteristics of intended receivers (Shatz & Gelman, 1973; Meissner & Apthorp, 1975). Yet under the difficult conditions of one-way communication to unobserved or unrevealing receivers, much older subjects behave as if they could not take the receiver's point of view into account (Flavell, 1976). It appears then, that egocentric writing in middle childhood and beyond does not result from an absolute inability to take another person's point of view. Rather, it would seem that egocentric writing arises from an incapacity to take account of the reader and cope with all the other demands of writing at the same time (Scardamalia, in press). Because the other (associative and performative) demands are ordinarily preemptive, social cognition simply gets left out.

Although immature writers may be aware of reader-related requirements, this awareness appears able to affect their writing only in gross ways. Scardamalia, Bereiter, and McDonald (in press) had students write instructions for playing a novel game. When alerted, by means of a demonstration videotape, to the possibilities for miscommunication of game rules, the effect on younger (grade 4) students was simply to induce them to say more and thus to cover more points, but not to express particular points more clearly. In another instruction-writing experiment (Bracewell, Scardamalia & Bereiter, 1978) reader saliency was varied from a condition in which no reader was specified to one in which the instructions were to be contained in a letter to a named unknown person, the letter sealed in an envelope, and addressed, with assurance that the letter would be delivered and a response returned. Subject of all ages, from 10 to adult, showed a tendency to adapt messages more to their readers when they expected feedback as to results; but not until grade 12 did they show a significant tendency to adapt messages to the supposed level of sophistication of their readers.

4. Unified writing. The defining characteristic of this next stage of writing is that it takes account not only of other people's perspectives but of the writer's own perspective as a reader. As readers, children do more than acquire knowledge about how writing should be done, which serves them in performative writing. They also grow in ability to respond critically and esthetically to what they read. These evaluative skills are not automatically applied to their own writing, however. Implicit in the present model is the speculation that it is easier to take the point of view of another person (as in communicative writing) than it is to take the point of view of oneself as reader.

Once one is able to integrate one's own evaluative reading skills with one's writing skills, however, an important feedback loop is established. Writing comes to be shaped according to one's own standards, which in turn are shaped by what one has written. As a result the writer begins to develop a personal style and a personal viewpoint. Writing becomes more authentic and satisfying. One does not merely write to entertain the reader but also to please oneself. One does not argue simply to convince the reader but also to present an argument that oneself finds convincing. From a rhetorical standpoint, however, the most important consequence of the establishment of this writing-reading feedback loop is that it leads to a focus on the written product as a thing to be fashioned. Hence, writing becomes a productive craft and not merely an instrumental skill.

Unified writing includes what Britton et al (1975) call poetic writing and which they characterize as creating "a verbal construct, an 'object' made out of language" (p. 90). As noted previously, however, their categories are genre-related, while the present scheme deals with cognitive systems. Although writing a well-crafted story and writing a logically coherent argument may differ in the particular rhetorical skills they draw upon, they would appear to be alike in the kind of processing mechanism involved; both require the involvement of one's own critical capacities in a writing-reading feedback loop.

5. Epistemic writing. Writing probably always plays an epistemic function in that our knowledge gets modified in the process of being written down. The satisfaction that many children derive from writing about their personal experiences probably does not derive only from sharing the experiences but also from what Church (1961) calls thematization--the processing of experience into verbal form. As the capacity for reflective thought develops, however, writing can take on a more complex epistemic function (Nystrand & Wiederspiel, 1977). Writing, because of its amenability to storage, review, and revision, makes possible a kind of extended and involved thought that is almost impossible without writing (Olson, 1977).

Epistemic writing emerges when the person's skill system for reflective thought is integrated with his skill system for unified writing. The creation of a literary work then becomes, in addition to whatever else it may be, a personal search for meaning. This seems to be what people mean by 'serious' literature, as contrasted with literature that exists solely as 'verbal artifact.' (It should be clear, however, that epistemic writing is not limited to literature of ideas. A search for meaning could be carried out entirely through narrative imagery, for instance, without any propositional thought at all.) The teacher-student 'process-conference' advocated by Graves (1978) appears to be aimed at helping students develop the epistemic potential of writing.

Epistemic writing represents the culmination of writing development, in that writing comes to be no longer merely a product of thought but becomes an integral part of thought. I becomes part of what Vygotsky has called the "extracortical organization of complex mental functions" (quoted in Luria, 1973, p. 31). As a result, the various skill systems that have been integrated to produce mature writing competence may now be enriched by the cognitive consequences of writing.

Writing Development and Schooling

The Schools Council research (Britton et al, 1975) illustrates the extent to which writing development, in a highly schooled society, is whatever the schools make it to be. Scribner & Cole (in press) show that in the rare case of the Vai, where a written language exists that is not mediated through schooling, writing is simply an adjunct to speaking and develops along with it. If writing were only what the schools make it, however, it is doubtful if anyone would get beyond the first two stages described in the preceding section. Traditional writing instruction, based on style manuals, models, and teacher correction, is almost wholly devoted to moving students from associative to performative writing. The alarm generated by the National Assessment of Educational Progress (1976) results and other less cogent findings has to do with the apparent failure of students to achieve performative writing, not with their failure to go beyond it.

Success in transcending or by-passing performative writing and moving students to higher stages is claimed for a number of innovative approaches, however. Several focus on communicative writing (Pettigrew, Tefft, & Van Nostrand, 1975; Elsasser & John-Steiner, 1977). The work of Kohl (1967) suggests that children who are not proficient at either performative or communicative writing may nevertheless undertake unified writing, in the sense of using writing to produce esthetically satisfying objects. Jacobs (1970) claimed success in moving elementary school children directly to epistemic writing appears also to be the goal of Elbow's (1973) unusual approach, designed for people who are 'hung up' on performative writing. His freewriting technique evidently induces regression to associative writing, following which a dialectical process of working out thoughts in writing is pursued.

These divergent experiments support the view that has already been suggested several times in this paper, that there is no natural order of writing development, in the sense of a fixed sequence which all writers must go through. At the same time, however, the experimental approaches suggest that it is at least meaningful to think of writing development in terms of discrete stages, each characterized by conscious focus on a particular aspect of writing. The stages may be ordered differently by different educational approaches, but they do not simply run together into amorphous growth.

It remains to be argued, however, that some orderings of stages may be more 'natural' than others. In order to mount such an argument, it is necessary to distinguish sharply between mastery and automaticity. If mastery of one kind of writing had to be achieved before progress to a higher-level kind of writing, then development would be impossible. Automaticity does not imply mastery but only proficiency such that the behavior in question requires little or no conscious attention. Examples of automaticity without mastery are to be found in people with fluent but illegible handwriting or with confident but inaccurate spelling. The preferred or 'natural' order of writing development would be a sequence of stages in which the attainment of automaticity at one stage maximally facilitates progress toward the next stage.

The sequence of stages depicted in Figure 1 would seem to be a 'natural' order on these terms. The attainment of fluency in associative writing frees attentional capacity for dealing with performative demands. (While concern with performative demands--that is, with meeting conventional standards of correctness in writing--may not be 'natural,' it is certainly consistent with the general readiness of children in the preadolescent years to be socialized to prevailing norms of rightness (Gesell & Ilg, 1946). Attaining a degree of mastery over stylistic conventions leads in turn to the discovery that writing can be used to affect the reader--that it can direct, inform, amuse, move emotionally, and so on. Thence emerges the communicative stage. Once students start writing for readers, it becomes a natural next step for them to start reading their own writing, which sets in motion the writing-reading feedback loop on which the stage of unified writing depends. Once this feedback loop is functioning well, it will be natural to discover that it leads not only to improved writing but also to improved understanding--that the loop constitutes a kind of dialogue with oneself. Thus may emerge the final, epistemic stage of writing development.

Indications for Research

One reason that speculations about writing development must be so tentative is that the available data mostly relate to what children do do under typical composition-writing circumstances and not to what they can do under maximally revealing circumstances. Chomsky (1964), noting the same condition in research on grammar in young children, pointed out the kinds of errors of interpretation that may result. Children may be wrongly judged to lack a certain competence simply because circumstances failed to elicit performance that would have revealed it. On the other hand, children may be wrongly judged to possess a certain competence on the basis of performance that is limited to the circumstances observed (e.g., a child might exhibit some formula expressions such as "He got run over by a truck" without having any general command of the passive transformation). Chomsky did not recommend any ways out of this predicament except broadly to suggest that "rather devious kinds of observations" would have to be made in different circumstances "so that a variety of evidence may be brought to bear on the attempt to determine what is in fact the underlying linguistic competence at each stage of development" (p. 36).

A parallel recommendation is clearly called for in writing research. Research based on 'writing samples' collected under conditions more-or-less resembling those of typical classroom writing have served to give us a picture of what written language and composition are normally like at different ages. With conclusion of the definitive studies of Britton et al (1975) and Loban (1976), it is likely that research of this kind has yielded about all that it can concerning the normal course of writing development. Further knowledge is more likely to come through the use of specialized tasks and other techniques that break through habitual school writing behaviors.

Footnotes

The writer wishes to thank Marlene Scardamalia for valuable suggestions and criticisms throughout the preparation of this paper.

[1] A T-unit is the minimal unit that can be punctuated as a sentence (whether the writer has punctuated it that way or not).

[2] In both the 1969 and the 1974 NACP testings, 17-year-olds scored slightly higher than 13-year-olds. If we compare the 17-year-olds of 1974 with the 13-year-olds of 1969, however, we find no gain, and these two groups represent approximately the same age cohort.

References

Adams, Marilyn Jager, & Collins, Allan. *A schema-theoretic view of reading comprehension*. ERIC Document No. ED 142 971, 1977.

Allen, Robert L. *English grammars and English grammar*. New York: Scribner's, 1972.

Blount, Nathan S. Research on teaching literature, language and composition. In Robert Morris William Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand-McNally, 1973.

Bobrow, Daniel G., & Norman, Donald A. Some principles of memory schemata. In Daniel Goreasko Bobrow & Allan M. Collins (Eds.), *Representation and understanding: Studies in cognitive science*. New York: Academic Press, 1975.

Bracewell, Robert J., Scardamalia, Marlene, & Bereiter, Carl. The development of audience awareness in writing. Resources in Education, in press.

Braddock, Richard, Lloyd-Jones, Richard, & Schoer, Lowell. Research in written composition. Champaign, Ill.: National Council of Teachers of English, 1963.

Britton, James, Burgess, Tony, Martin, Nancy, McLeod, Alex, & Rosen, Harold. The development of writing abilities (11-18). London: Macmillan Education, Ltd., 1975.

Case, Robbie. Gearing the demands of instruction to the developmental capacities of the learner. Review of Educational Research, 1975, 45(1), 59-87.

Case, Robbie. Implications of developmental psychology for the design of effective instruction. In Alan M. Lesgold, James W. Pellegrino, Sibke D. Fokemma, & Robert Glaser (Eds.), Cognitive psychology and instruction. Plenum, New York: Division of Plenum Publishing Corporation, 1978.

Chomsky, Noam. Formal discussion. In Wick Miller & Susan Ervin (Eds.), The development of grammar in child language. Society for Research in Child Development Monographs, 1964, 29(1), 35-39.

Church, Joseph. Language and the discovery of reality. New York: Random House, 1961.

Collis, Kevin F., & Biggs, John B. Classroom examples of cognitive development phenomena. ERDC Funded Project 7/41, University of Newcastle, undated.

Diebel, Amelia, & Sears, Isabel. A study of common mistakes in pupils' written English. Elementary School Journal, 1917, 18, 172-185.

Elbow, Peter. Writing without teachers. London: Oxford University Press, 1973.

Elsasser, Nan, & John-Steiner, Vera P. An interactionist approach to advancing literacy. Harvard Educational Review, 1977, 47, 355-369.

Emig, Janet. The composing processes of twelfth graders. Champaign, Ill.: National Council of Teachers of English, 1971. (Research Report No. 13.)

Flavell, John H. Cognitive development. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Flower, Linda. Writer-based prose: A cognitive basis for problems in writing. Unpublished paper, Carnegie-Mellon University, Pittsburgh, Pennsylvania, undated.

Fodor, Jerry A., Bever, Thomas G., & Garrett, Merrill F. The psychology of language: An introduction to psycholinguistics and generative grammar. New York: McGraw-Hill, 1974.

Gesell, Arnold Lucius, Ilg, Frances L., & Ames, Louise Bates. The child from five to ten. New York: Harper and Bros., 1946.

Gleason, Henry Allan, Jr. Linguistics and English grammar. New York: Holt, Rinehart & Winston, 1965.

Goody, Jack, & Watt, Ian. The consequences of literacy. Comparative Studies in Society and History. 1963, 5, 304-345.

Graves, Donald H. Children's writing: Research directions and hypotheses based upon an examination of the writing process of seven-year-old children. Unpublished doctoral dissertation, State University of New York at Buffalo, 1973.

Graves, Donald H. *Balance the basics: Let them write*. New York: Ford Foundation, 1978.

Havelock, Eric Alfred. Prologue to Greek literacy. In Cedric Boulter (Ed.), Lectures in memory of Louise Taft Semple, second series. Cincinatti: University of Oklahoma press for the University of Cincinatti, 1973.

Hayes, John R. Identifying the organization of writing processes. Paper presented at Symposium on Cognitive Processes in Composition, Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 22, 1978.

Hill, John David. An analysis of the writing of elementary children, grades 2 through 6, to determine the presence, frequency of use and development by grade level of specified literary devices. Ed.D. thesis, Indiana University, 1972. Order # 73-10, 770.

Hunt, J. McVicker. Piaget's observations as a source of hypotheses concerning motivation. *Merrill-Palmer Quarterly of Behavior and Development*, 1963, 9(4), 263-276.

Jacobs, Gabriel H.L. *When children think*. New York: Teachers College Press, Columbia University, 1970.

Keeney, Mary Louise. An investigation of what intermediate-grade children say about the writing of stories. Ed.D. thesis, Lehigh University, 1975. Order # 76-5091, 127 pp.

Klahr, David, & Wallace, John Gilbert. *Cognitive development: An information-processing view*. Hillsdale, N.J.: Lawrence Erlbaum, 1976.

Kohl, Herbert. *Thirty-six children*. New York: Signet Books, 1967.

Loban, Walter. *The language of elementary school children*. Urbana, Ill.: National Council of Teachers of English, 1963. (Research Report No. 1.)

Loban, Walter. *Problems in oral English*. Urbana, Ill.: National Council of Teachers of English, 1966. (Research Report No. 5.)

Loban, Walter. *Language development: Kindergarten through Grade Twelve*. Urbana, Ill.: National Council of Teachers of English, 1976. (Research Report No. 18.)

Long, Ralph Bernard. *The sentence and its parts: A grammar of contemporary English*. Chicago: University of Chicago Press, 1961.

Looft, William R. Egocentrism and social interaction across the life span. *Psychological Bulletin*, 1972, *78*(2), 73-92.

Luria, Aleksandr Romanovich. *The working brain: An introduction to neuropsychology*. London: Penguin Books, 1973.

Lyman, Rollo La Verne. Summary of investigations relating to grammar, language and composition. Chicago: University of Chicago, 1929. (Supplementary educational Monographs, No. 36, published in conjunction with *The School Review* and *The Elementary School Journal*.)

Meissner, Judith A., & Apthorp, Helen. Nonegocentrism and communication mode switching in black preschool children. *Developmental Psychology*, 1976, *12*(3), 245-249.

Mennig-Peterson, Carole, & McCabe, Allyssa. Structure of children's narratives. Paper presented at the biennial meetings of the Society for Research in Child Development, New Orleans, March, 1977.

Moscovici, Serge, & Plon, Michel. Les situations-colloques: Observations theoriques et experimentales. *Bulletin de Psychologie*, 1966, *19*, 702-722.

National Assessment of Education Progress. Writing mechanics 1966-1974: A capsule description of changes in writing mechanics. Report No. 05.W-01. Denver, Colorado: National Assessment of Educational Progress, 1975.

Nold, Ellen W. Revising. In Carl H. Frederiksen, M.S. Whiteman & J.F. Dominic (Eds.), Writing: The nature, development, and teaching of written communication, in press.

Nystrand, Martin. Assessing written communicative competence: A textual cognition model. Toronto, Canada: The Ontario Institute for Studies in Education, 1977. (ERIC Document Reproduction Service No. ED 133 732)

Nystrand, Martin, & Widerspiel, Michael. Case study of a personal journal: Notes towards an epistemology of writing. In Martin Nystrand (Ed.), Language as a way of knowing: A book of readings. Toronto: The Ontario Institute for Studies in Education, 1977.

Odom, Richard D. A perceptual-salience account of decalage relations and developmental change. In Charles J. Brainerd & Linda S. Siegel (Eds.), Alternatives to Piaget: Critical essays on the theory. New York, Academic Press, 1978.

Odom, Richard D., Cunningham, Joseph G., & Astor, Eileen C. Adults thinking the way we think children think, but children don't always think that way: A study of perceptual salience and problem solving. Bulletin of the Psychonomic Society, 1975, 6, 545-548.

O'Donnell, Roy C., Griffin, William J., & Norris, Raymond C. Syntax of kindergarten and elementary school children: A transformational analysis. Champaign, Ill.: National Council of Teachers of English, 1967. (Research Report No. 8.)

Olson, David R. From utterance to text: The bias of language in speech and writing. Harvard Educational Review, in press.

Pascual-Leone, Juan. A mathematical model for the transition rule in Piaget's developmental stages. Acta Psychologica, 1970, 63, 301-345.

Pascual-Leone, Juan, & Goodman, Doba. Intelligence and experience: A neo-Piagetian approach. Unpublished manuscript, York University, 1976.

Pascual-Leone, Juan, & Smith, June. The encoding and decoding of symbols by children: A new experimental paradigm and neo-Piagetian model. Journal of Experimental Child Psychology, 1969, 8, 328-355.

Pettigrew, Joan, Tefft, Nancy E., & Van Nostrand, A.D. Functional writing: Secondary school edition. Providence, R.I.: Center for Research in Writing, 1977.

Posner, Michael, & Keele, Steven W. Skill learning. In Robert Morris William Travers (Ed.), Second handbook of research on teaching. Chicago, Rand-McNally, 1973.

Scardamalia, Marlene. Two formal operational tasks: A quantitative neo-Piagetian and task analysis model for investigating sources of task difficulty. In Gerald I. Lubin, James F. Magary, & Marie K. Poulsen (Eds.), Piagetian theory and its implications for the helping professions (Eds.), Piagetian theory and it implications for the helping professions. Los Angeles: University of Southern California Publications, 1975.

Scardamalia, Marlene. How children cope with the cognitive demands of writing. In Carl H. Frederiksen, M.S. Whiteman, & J.F. Dominic (Eds.), *Writing: The nature, development, and teaching of written communication*, in press.

Scardamalia, Marlene, Bereiter, Carl, & McDonald, James D.S. Role-taking investigated by manipulating anticipatory knowledge. *Resources in Education*, in press.

Schaeffer, Benson. Skill integration during cognitive development. In Alan Kennedy & Alan Wilkes (Eds.), *Studies in long term memory*. London: John Wiley & Sons, 1975.

Scribner, Sylvia, & Cole, Michael. Unpackaging literacy. In Carl H. Frederiksen, M.S. Whiteman, & J.F. Dominic (Eds.), *Writing: The nature, development, and teaching of written communication*, in press.

Shantz, Carolyn Uhlinger. The development of social cognition. In E. Mavis Hetherington (Ed.), *Review of child development research* (Vol. 5). Chicago: University of Chicago Press, 1976.

Shatz, Marilyn, & Gelman, Rochel. The development of communication skills: Modifications in the speech of young children as a function of the listener. *Monographs of the Society for Research in Child Development*, 1973, *38*, (5, Serial No. 152).

Shaugnessy, Myna P. *Errors and expectations: A guide for the teacher of basic writing*. New York: Oxford University Press, 1977.

Shuy, Roger W. Toward a developmental theory of writing: Tapping and knowing. Paper presented at National Institute of Education Conference on Writing, Los Alamitos, California, June, 1977.

Simon, Herbert A. On the development of the processor. In Sylvia Farnham-Diggory (Ed.), *Information processing in children*. New York: Academic Press, 1972.

Simon, Herbert A. How big is a chunk? *Science*, 1974, *183*, 482-488.

Simon, Jean. *La langue ecrite de l'enfant*. Paris: Presses Universitaires de France, 1973.

Stahl, Abraham. Structural analysis of children's compositions. *Research in the Teaching of English*, 1974, *8*, 184-205.

Stahl, Abraham. The structure of children's compositions: Developmental and ethnic differences. *Research in the Teaching of English*, 1977, *11*, 156-163.

Stallard, Charles K. An analysis of the writing behavior of good student writers. Research in the Teaching of English, 1974, *8*, 206-218.

West, William W. Written composition. *Review of Educational Research*, 1967, *37*(2), 159-167.

A SCHOOL PERSPECTIVE ON LANGUAGE: A DISCUSSION OF READING AND WRITING

Parker Damon
Principal
McCarthy-Towne School, Acton, MA

Abstract

It is artificial to separate the discussion of reading and writing from speaking and listening. This point is just as true for early adolescents as it is for students in the primary grades. Expectations for reading instruction and assessment should be considered in terms of the answers to three questions: Do we know how to teach reading? Is learning to read difficult? and Is there a reading problem? The author's answers to the questions are, respectively, "yes," "no," and "only for some children." The lack of enthusiasm early adolescents feel for reading, writing, and school in general is a more crucial problem. Therefore, the teaching of reading and writing should build upon the developmental phases of this age group by providing a more integrated language arts curriculum that contains more purposeful activity. Determination of the curriculum's success should be by assessing the students' involvement in a range of activities as much as by relying on test scores. The NIE should support and promote changes in public, parental, and professional attitudes away from a reliance on tests and towards other assessment techniques. The NIE should support the inservice needs of teachers by making more available to professionals and the public information about what works, what should be tried, and what specific techniques and technology are needed. The teaching of writing is in particular need of NIE support since there is much confusion among teachers about what to do to teach writing better. Writing instruction, similar to that of reading, is fragmented by curriculum materials and tests, and is separated from the curriculum for speaking and listening. The pedagogical danger is for teachers to get in the way. Often learning to read is made more difficult than necessary, and becoming a better writer is made cumbersome with excessive teacher involvement. The objective for teachers is for them to subordinate their teaching to the students' learning.

INTRODUCTION

The discussion which follows touches on four topics that should be treated as one: the teaching and learning of reading and writing. In the elementary school, this part of the curriculum is labelled language arts in an effort to integrate all aspects of communication--listening, speaking, reading, and writing. Therefore, there are really eight topics which should be considered simultaneously and with regard to what students do. None of these activities occurs in a vacuum; each is linked to some form of doing as well as being part of the other three. However, as students progress from grade to grade these relationships tend to be forgotten. Speaking, listening, and doing drop away as educators attend to reading and writing. Though this paper is caught in the same trap, its main aim is to suggest some ways out.

READING

> Red will read what Reid read about reeds.
> Then the two will be able to:
>
> a) understand each other, b) weave similar baskets, c) play music better, d) measure their rugs, e) use a bow, f) space warp yarns evenly.

What is it we expect from the teaching of reading, and how do we know when the job has been done? Before I try to answer this two-part question, I would like to spend some time answering several others: do we know how to teach reading to young non-readers? Is learning to read difficult? Is there a reading problem?

Do we know how to teach reading to young non-readers?

Yes. At least we do in our school for all except 3 percent of them. These unfortunate students do not seem to improve no matter what is done for them. Sometimes their difficulties appear to be a collection of identifiable and usually complex emotional, social, and physical problems. Interspersed with these students are others with hearing and speech problems, language processing and production problems, and visual and motor problems. I am not referring to the more severely mentally and physically handicapped students infrequently found in the classroom.

How much of a school's time and resources should go toward helping these students? I can identify, in our school, about seven students (1 percent) who might be labelled as serious non-readers. These would be primary grade students who have yet to break the decoding barrier, and intermediate students who can decode but can't comprehend anything but the simplest sentences and paragraphs. Added to them are fifteen or so more who have a mixture of language, math, and other schooling problems for a total of about 25 students who receive extra help from one full-time professional faculty member and one part-time semi-professional. Altogether, the school allocates about $25,000 toward helping these students. Should we spend more? Are we spending too much in terms of what can be expected, or not wisely enough in keeping with what is known? If every school were the same as ours, then the country would be spending about 840 million dollars per year to assist the 840,000 students in grades K-7 who are non-readers. If the number of students sharing this problem is larger nationwide, and the percentage of money being spent is proportionately less, then is our school, or the country, spending too much, too little, or too unwisely?

Is it possible to equate the teaching of reading to the spending of money? Is there a bottom line to the teaching of reading?

Is learning to read difficult?

No. As suggested by the answer to the preceding question, most young children learn to read. Many already know before they come to school and about 50 percent learn to do so by the time they are finished second grade. (In our school the majority can read by midyear of first grade.) These children appear to learn to read regardless of what we do, or in spite of what we do. The point we try to remember is to stay out of their way. By the end of third grade, three quarters or more of our students are reading. They are reading with understanding, are able to do so for a variety of purposes, and usually like to read. These students present teachers and administrators with the dilemma of finding enough time within the school schedule for reading and reading-related activities, and the problem of providing sufficient quality and quantity of books for students to read.

Starting in grade one, but beginning to peak in grade three, is professional and parental concern for the progress of the children taking longer to learn to read. The debates about code vs. meaning emphasis, aural vs. visual learning, individualized vs. grouped instruction, primers vs. trade books, and the like, are focused on this 25 percent or so of the students in the early grades. Again, in our school we have to remember to keep out of the way. That is, we have to keep from making the task more difficult than it really is by becoming impatient, by applying too much pressure, by not "subordinating our teaching to the students' learning."[1] Undoubtedly, for some of these students some materials and approaches are better than others, and many teachers do need and want more support in providing help for these children.

This support can be of different types. First of all, there needs to be more public, parental, and professional acceptance of the fact that learning to read occurs within a latitude of time, that the course and duration of the journey are not as important as the common destination. As long as there is a lack of acceptance of this fact, those teachers and principals who want to provide students with a wide latitude will be prevented from doing so by the pressures of this non-acceptance. Second, many teachers seek and welcome direct help from other profesionals when particular students appear to be foundering. But this extra consultant help has to be available, and teachers have to have the time to take advantage of it. A third type of support is in-service. Many teachers, specialists, and principals know they do not know enough about what surrounds and is incorporated in the reading experience. They are well aware, however, that a day or afternoon a month, or a week or so a summer is not going to provide them with long-term, day-to-day, applicable benefits.

---

[1] I am indebted to Caleb Gattegno's work. See for example his book WHAT WE OWE CHILDREN: The Subordination of Teaching to Learning, Avon Books, 1970.

What to do for and with students who take longer to learn to read does require attention, but of a type that goes beyond an examination, dissemination, and implementation of "better" materials and methods. Professionals inside elementary schools (and I believe secondary schools also) need to continually work on ways to keep learning to read from becoming difficult. The research and practice of the past fifty years or so, although well intentioned, seem to have been going in the opposite direction. For example, educators seem to pay more attention to what they think should be done and less to what is indicated by what children say and do.

Children come to school knowing how to talk, and wanting to do so. They like to play games, be active, make-believe, explore, and find out about things. Existing instruments for assisting and assessing language learning, particularly reading, tend to fragment instruction, and neglect important areas such as oral communication. These materials and methods make all too little use of what children like to do, want to do, and could do. My colleagues at the school where I work know how important yet difficult it is to not get in the way, to utilize interests and abilities, and to provide meaningful and interrelated learning opportunities. Worksheets and tests that focus on skills and drills rarely do this. We, in our school, have found that games, group tasks, physical activity, opportunity for individual inquiry and reflection, varied forms of self-expressions, and the chance to take risks are the main components of the children's curriculum. Our job is to try to join theirs with ours.

Is there a reading problem?

Yes, for some children; learning to read for most children is not difficult. I think there are other problems which tend to be reflected as reading problems, but what to do about the reading is considered more easily handled through research, design, trials, and large-scale restructuring. Schools are not receiving much assistance in changing themselves as family living and work patterns shift and as the economy, values, and culture of the community alter. The length of the school year and day remains the same. The schedule and content of what goes on inside the school, as well as the little bit outside, also remain the same.

For example, at our school we have noticed the following changes in our entering kindergarteners during the past five years. Perhaps 25 percent or so exhibit these traits either strongly in isolation or strongly when considered collectively. Members of this 25 percent are not as ready for group activities, but more interested in pursuing individual play such as block building. They are not as good listeners, at least in group settings; it seems more difficult for them to pay attention to instructions and to follow directions, to listen to stories as a member of the whole class. They are not as communicative with others around them, be they children or adults. They are not as self-confident and outgoing as their classmates and tend instead to either hold back and not participate or be overactive and disruptive. As a result, what we have come to expect of ourselves and these students has had no change. In kindergarten, we try to give these students more individual adult attention (not the non-adult, attentionless placebo of individualized instruction), more physical and social activity, more time to become school-oriented and able to accept schooling. What we try to provide is some latitude for these children to develop as they might have been able to had they had more time at home and different kinds of experiences during that time.

It is self-evident to us that these changes and the need for this latitude are the by-products of our community's growing number of single-parent families and families in which both parents work. An increasing number of our students attend some form of day-care program before and or after school or are alone at home with a slightly older sibling for these periods of time. Some of these children seem unaffected by their irregular schedule and independent existence. Unfortunately, a sizable number do seem affected socially, emotionally, and in the ability to receive and produce language activity. We believe these are the children for whom there has not been sufficient talking with, listening to, reading aloud beside, and writing for. These are the children for whom learning to read is a problem, and solving this problem is much more complex than we sometimes admit. The social theory of reading problems is not new, but it is often forgotten as we look for new strategies and supplies for teaching poor readers.

What is it we expect from the teaching of reading, and how do we know when the job has been done?

My school system asked itself the first part of this questions several years ago and some of us came up with this answer:

> Is it not to open up the world of books so that the child can know a larger world than he himself can experience? so that he can gain independence in gathering information? so that he can have a medium to engage pleasurable experiences and excite his imagination? If this is our purpose as teachers of children, can we then separate the teaching of reading skills from the world of books? Is literature not an integral part of the reading program?

This part of the answer goes beyond those concerned with minimum performance abilities of reading labels, directions, and instructional material in another subject area or grade level. Our answer continued with the following thoughts:

> Meaningful experiences with literature are a way of knowing. A child's enjoyment of, involvement with, and response to the literary experience makes possible discoveries about himself

> and his world. It is the teacher's role then to ensure children opportunities for enjoyment, involvement and response to literature.[1]

Although, as the footnote indicates, this response was intended to apply to the teaching of reading to students in the intermediate grades (4 - 6 in our system), I believe it provides a more universal focus.

In our school, for example, the teaching of reading in the primary grades relies as much, if not more, on trade books for the source of student reading matter as it does on teacher-made work sheets, commercial workbook materials, primers, and other reading textbooks. One reason for the system's wanting an answer to the question was to legitimize this emphasis both in our school and system-wide. Another reason, however, was to begin addressing a related area of concern: that students in the system's intermediate grades and on up into junior high school were exhibiting less enthusiasm for reading and schooling. The issue was not just that many students seemed to be reading poorly and needed remedial help. There was also the awareness that much of what we were doing did not make pedagogical sense to ourselves and purposeful sense to our students. We were, and are, concerned with student boredom and disinterest brought about in part by schooling and in part by the many changes occurring in early adolescents.

The school experiences of early adolescents, the majority of whom know how to read, should be the quintain of our labor to improve instruction in any particular subject area. Whether it be the study of literature, history, or art, or the development of skills such as reading, writing, or computing, researchers and practitioners alike should pay more attention to what early adolescents think about, are interested in, want to try, and actually do. The need for adults in schools to pay attention is probably greater during this period of growth than earlier in that early adolescent children are less forgiving than their younger schoolmates. But there are more serious reasons. The pervasive excitement and luster of schooling as a new experience is wearing off; the conflict in shifting from student centeredness to subject centeredness is beginning; the pressure for students to be prepared for something are mounting ("time is running out," "is she doing as well as he is?" "what are they missing?" are phrases more often hear as students get older).

The teaching of reading to early adolescents must take into account the many issues surrounding their phase of development. To know whether or not the expectations for the teaching of reading to early adolescents are being met means to know not only how students perform, but also what schools permit. I believe many schools are not permitting or encouraging enough that is new, challenging, interesting, and helpful to students in the middle school years. Much, if not most, of the curriculum for students age 9 to 13 fails to live up to these criteria on a year-to-year basis.

---

1 Using Children's Books as Part of the Reading Program in the Intermediate Grades. A Research and Development Project, Summer 1975. Acton Public Schools, Acton, MA, vi.

This is particularly so for the language arts curriculum in which there is not enough doing or purposeful activity.[2] Schooling during these years ia a holding pattern.

For example, how much of what goes on is a necessary follow-up of earlier work in school, or how much is a necessary prerequisite for work to be done later, or what is introduced that is totally new that may be followed-up on if the student wants to? Turn the question around. Why is there so much repetition? so little enthusiasm by students for school work? The enthusiasm is there for other things so the answer rests with more than "it's the nature of students this age."

For example, our school and then the system introduced an Outward Bound ropes training types of program for elementary students.[3] The program provides students with real physical, emotional, and social risks. While challenging them it encourages them to grow in self-confidence, coordination, cooperation, independence, and respect for others. Although all students in our school participate, the program works best with students in grades 4 - 8. They find it exhilarating. Our desire is to make the rest of their schooling be this way too. Fortunately, there are more and more educators around the country who recognize this need and share the goal. One place to begin is with more active uses of language in the elementary school, particularly the intermediate grades. As already noted, the National Association of Elementary School Principals has begun coordinating some of the work to improve the curriculum experiences for early adolescents.[4] Individual teachers, isolated schools, and even independent and innovative districts by themselves cannot bring about meaningful and lasting change in curriculum and schooling practices. The educational complex is just too big. If the expectation for the teaching of reading, or any other part of the curriculum, is to be similar to the goal just described, then a broader national direction and union of professionals are required.

The second part of the question, "how do we know when the job has been done?" is more difficult to answer. In our school, we try to ask ourselves this subset of questions about our teaching of reading. Is there evidence that our efforts have:

-- opened up the world of books so that the child can know a larger world than he himself can experience?

-- provided the opportunity for each student to gain independence in gathering information?

---

[2] See Active Use of Language: A Program to Improve Literacy Skills for Elementary School Children, a Planning Proposal Submitted to the U.S. Office of Education, June 15, 1978, by the National Association of Elementary School Principals.

[3] See Trip, Caroline E. Project-Challenge. Acton-Boxborough Regional School District, Acton, MA, 1977.

[4] See "Listening, Speaking, Reading, Writing, Doing," The National Elementary Principal, No. 4 (June 1978) Vol. 57.

-- provided a medium for engaging in pleasurable experiences and exciting the imagination?[5]

But we find this form of assessment difficult to conduct. Subjective judgments about how, what, when, and why a student reads are not as trusted as objective, standardized, norm-referenced tests purporting to indicate how well a student, or group of students, reads, or to compare reading abilities of different students. The reliance on standardized tests is unfortunate for it leads to all sorts of waste.[6] I agree wholeheartedly with Harold Howe's address at the National Conference on Achievement Testing last March. In particular, the following comments seem most pertinent:

> NIE should support major research on what are sometimes called the "soft" factors in judging students. These are the non-quantifiable aspects of human beings that must be considered in decisions about them.
>
> ... NIE can help to work on the serious problem of public misunderstanding of tests.
>
> ... [there is a need for NIE to help in the] improvement of testing that really helps teachers in the diagnosing.[7]

Thus, in order to know when the job has been done, two developments are necessary. First, exclusive dependency on the forms of tests has to be reduced. Standardized norm or criterion reference tests do not help answer questions such as those I have asked above. Nor do they really tell us much about how well a student reads in isolation or in comparison with others. Second, better tests need to be created. The design criteria for such tests should include these requisites:

-- not be secret
-- not rely on memory
-- be self-correcting
-- use constructed answers
-- have several items per probe
-- indicate possible reasons for errors and alternative acceptable answers
-- not be timed
-- allow different methods of administration
-- be unthreatening (similar to usual school experiences)

---

[5] Op. cit. vii.

[6] See Damon, J. Parker, "Questions You Should Ask About Your Testing Program," in The Myth of Measurability, Paul L. Houts, ed. New York: Hart Publishing Company, Inc. 1977.

[7] Howe, Harold, "Tests and Schooling." An Address at the National Conference on Achievement Testing, Washington, D.C., March 1-3, 1978.

A model for this type of test exists for mathematics.[8] There is every reason to believe the design of a similar kind of test for reading is also possible. It may also be possible to fashion an array of language arts tests which not only diagnose or "improve the capacities of teachers to tune their teaching methods and materials to the needs of the individual child,"[9] but also serve to help interrelate reading, writing, speaking, and listening. Currently, tests separate these components of the language arts curriculum, and fail to consider a significant aspect of all language use: situation.

Reading tests, for example, do not account for the fact that people read differently at different times, in different places, and with different people. When a child is reading out loud to a sibling at home it is not the same as reading to a parent, nor to a friend at school, nor before the entire class, nor alone with the teacher. Reading to answer someone else's questions is not the same as reading to answer one's own or those questions contained as part of the material being read (and these may be either the explicit, "did you know that?" type or the implicit variety, "will the child get rescued?").

The reading situation is as important as the content of what is being read. Reading tests give too much emphasis to content and not enough to situation, and this emphasis is reinforced by the authority of reading formulae, graded word lists, and leveled reading materials. In an article appropriately entitled "I Never Wrote for Children," P.L. Travers, the author of Mary Poppins, refers to C.S. Lewis's statement "that a book that is written solely for children is by definition a bad book."[10]

WRITING

> Human language is shaped by the minds of infants. Speaking is innate; writing is an invention.[11]
>
> I did not so much gain the knowledge of things by the words as words by the knowledge I had of things.[12]

---

[8] See the mathematics computation tests developed by Project TORQUE, Education Development Center, Newton, MA.

[9] Howe, Harold, op. cit., 15.

[10] Travers, P.L., "I Never Wrote for Children," The New York Times Magazine, July 2, 1978, 17.

[11] Miller, George. Maxim Number 6, Presented at the National Association of Elementary School Principals Conference on the Active Uses of Language, Education Development Center, Newton, MA, June 5-7, 1978.

[12] Plutarch from Eduard C. Lindeman's Introduction to Life Stories of Men Who Shaped History from Plutarch's Lives. New York: A Mentor Book, 1950.

Do young children like to write?

Young children like to hear stories, see pictures that tell stories, tell stories, draw pictures that tell stories, have someone else write down stories they tell, be told how to spell words so they can write the stories themselves, talk about the stories they have drawn or written, read the stories they or someone else has written down. A significant number of children are more interested in writing than in reading. Some of these children will spend from thirty minutes to an hour writing their name and other words they know over and over. Sometimes they can read what they have written aloud, sometimes not. Most primary grade level teachers do not have time to pursue with all students these kinds of activities on a one-to-one or even one-to-three basis. Many of these teachers acknowledge that they need help in extending and building upon students' interests in writing. Others who know how to further develop their students' interests and abilities find themselves restricted by the formalities of a syllabus. Perhaps for these reasons students begin to lose their interest in writing as they proceed from grade to grade. Perhaps it is the drudgery of filling in countles words and phrases on work sheets, or seeing their biweekly composition only corrected for proper margins, neat handwriting, correct spelling and mechanics. Perhaps it is because no one reads and responds to what they have written.

Is there a writing problem?

Unlike the position I have taken on reading--that there is not a widespread reading problem--I do believe there is a writing problem in most schools for most students. The previous paragraph provides some of the reasons why I think the problem exists. In addition, there are these influences:

-- teachers find themselves caught in the communication loop of having to respond to most everything each student writes. The word "respond" usually means correct or proofread rather than to ask questions, critique passages, or detail suggestions. If they do not respond, and quickly, students, parents, and supervisors get angry.

-- technology that could help with the teaching and learning of writing has either not been developed or not utilized. Typewriters have been around for a long time but are still by and large a classroom oddity. Tape recorders, and their relative the language laboratory, are infrequently used as a part of the writing process. Computer terminals and television display screens are barely at the experimental classroom stage.

-- tests, both teacher and commercially made, do not value writing. All teachers feel the pressures of time when caught in the bind of either being able to assess what many students know or assisting a few to state what they know more fluently. Existing instruments for assessing language learning tend to fragment instruction, and neglect important areas such as oral and written communication."[13]

-- topics, activities, and involvements of various kinds that are important to the students themselves need to be the subject matter of their writing. This point is particularly significant for students age 9 to 13. Workbooks, texts, teacher's guides, and supplementary materials do not help teachers find out what interests particular students, and then use writing to build upon these interests. Students should be able to answer for themselves the question of why what they write is important.

These are a few of the influences contributing to the writing problem; they also suggest ways to solve the problem.

I think there is a problem because I have worked with students of all ages who have said and shown that they would like to be able to write better. I have talked with a lot of teachers who would like to do a better job of teaching writing. I have heard numerous principals and language arts supervisors complain that they were not able to give their teachers adequate support to teach writing well. And I have listened to parents and other citizens decry the quality and quantity of writing done in school. I think this information is a more accurate indicator of the problem than the data supplied from standardized tests. In fact, some tests such as the National Assessment indicate improvement in writing ability. As far as I can determine, there is no correlation between multiple choice assessments of writing ability, how students perform in the classroom, and what they do on their own. Marginal correlation exists between tests containing writing samples and students' other written work. The "soft" factors in judging students that Harold Howe referred to are most crucial where writing is concerned.

Three more suggestions for the teaching of writing

Much of what is needed to improve the teaching of writing has already been discussed. It should also be remembered that these improvements cannot occur as isolated exercises; the use of language is common to all subjects. Writing will improve when it is used more. The teacher's responsibility is to make writing enjoyable, involving, responsive (i.e., it should be part of a meaningful sequence of communication between the student and others), purposeful and, for early adolescents, capaitalize on their enthusiasm. I believe that there are ways for teachers to meet this responsibility. The next three suggestions outline how this task might be accomplished.

---

[13] Active Use of Language: A Program to Improve Literacy Skills for Elementary School Children. op. cit., p. 1.6.

First, the model for teaching writing which makes the most sense to me is described by Moffett in Teaching the Universe of Discourse.[14] This book and the other materials supporting the Interaction curriculum provide an excellent basis for both theoretical and practical development of students' writing, reading, speaking and listening abilities.

Second, the mode for teaching writing should make use of every piece of available technology. Typewriters, tape recorders, computer terminals, and television display screens can make writing easier, more active, exciting, and instructive. These tools can not only broaden each student's communication network, but also release the teacher from the neessity or obligation of participating in each one.

Third, the module for determining whether or not writing instruction is successful will be the work students produce school-wide rather than work done on tests, exercise sheets, and isolated compositions. As with reading, teachers will need support from outside the school as their points of view, emphasis, and technique shift from relying on test results to samples of student work. The NIE could provide this kind of public support. In addition, the NIE could help school districts explain to the public why the need for teacher inservice programs is required if improving the teaching of writing is expected.

A closing comment

Much can be done to improve the quality of writing instruction. As with reading, the question is not what to do but when to begin and with whose leadership. There are numerous examples of experiences which consistently help students improve their writing ability. The problem is to get these examples of what works into widespread use. The NIE could make what is known about the teaching and learning of writing (or any of the language arts) more accessible and acceptable. How much time should teachers and students devote to writing, and do teachers have the time to make this an effective experience? Do the emphases on creative writing in the primary grades and expository writing in the intermediate grades have to be placed in opposition to each other? Can the improvement of students' abilities to listen, speak, read, and write be subsumed under the activities and interests students find exhilarating?

Our school, on its own initiative but with the support of parents and school system, used to reallocate 10 percent of its total operating budget for inservice programs designed to help us put into practice the answers to questions such as these. The money ($50,000) came from not having some staff positions found in the other district schools. The faculty was willing and able to take on additional responsibilities as well as face the challenge of a different pedagogy because there was outside support. As a result, the faculty began to find ways to maintain and increase students' excitement for many areas of the curriculum and, as a by-product, improve students' ability to use language. But the faculty has become increasingly unwilling to devote their time and energy to this kind of inservice commitment because outside support has shifted from improvement of instruction to the equalization and expansion of other school programs.

---

[14] Moffett, James. Teaching the Universe of Discourse. Boston, Houghton Mifflin Company, 1968; see particularly page 47.

By providing both the public and professionals in schools with information designed for immediate use, the NIE and other government agencies could provide the leadership and outside support for a refocus on the improvement of instruction: on the difficult problem of subordinating the teaching of reading and writing to the learning of listening, speaking, reading, and writing for a purpose.

AUTOMATED DICTIONARIES*

George A. Miller
Experimental Psychology Laboratory
Rockefeller University

A dictionary is a treasury of information about a language. It contains information about spelling, pronunciation, etymology, syntactic category, meanings, accepted usages, even pictures, for thousands and thousands of different words. It is valuable for both writers and readers. It is also valuable to teachers and students, because it is a codified representation of a large body of linguistic knowledge that literate adults are expected to have.

Nevertheless, most adults would probaby agree that they do not use dictionaries as often as they should. One reason is that referring to a dictionary interrupts the flow of thought. It takes time, first to find the dictionary, then to find the word, then to select the desired information from the lexical entry. Few adults -- and even fewer children -- have the sustained attention span required to bridge such interruptions.

If there were sufficient value in the project, the automation of a dictionary would be well within the power of existing technology. It would be possible, for example, to type a word into a computer keyboard and have the machine pronounce it. Any information stored with that word could be displayed on a CRT almost instantaneously. The tedious task of searching through an alphabetical listing is one that machines can perform much faster than humans. A central computer could easily serve a large number of classrooms. Let us refer to such a system as Automated Dictionary Model 1, or AD-1 for short.

Let us imagine that AD-1 is available and consider whether any interesting research projects might result. The most obvious question to ask is whether anyone would use it, or whether it would be as neglected as its printed progenitor. User studies might suggest ways to make the system easier to refer to, or more accessible to more people. It is likely, however, that the system would be ignored unless special efforts were made to educate people in its value, or special tasks were devised to use the system in integral ways.

A second line of research, therefore, would be to develop interesting and educational tasks in which AD-1 played a useful role. Some examples:

-- Spelling. A word is pronounced and the student types it into the keyboard. If the spelling is correct, AD-1 confirms it; if not, various responses might occur, ranging from nonrecognition to corrected spelling suggestions (based on tabulations of the most frequent spelling mistakes).

---

*This paper was prepared and circulated during the course of the conference.

-- Word recognition. Letter sequences could be presented and the student asked whether or not they spell an English word. AD-1 could serve as referee in student competitions.

-- Anagrams. The student could be challenged to form as many words as possible from a given collection of letters, and given access to AD-1 to settle dubious instances. (Scrabble, word pyramids, etc.)

-- Word search. The dictionary definition of some sense of a word could be given and the student challenged to think of the word. Guesses would be typed into AD-1, which would display the definitions of that word, until a matching definition was found. The definitions might be provided in the form of clues for the solution of crossword puzzles.

-- Picture naming. A picture from the dictionary might be displayed and the student challenged (as in word search) to think of its name.

These and other dictionary games could be explored for their intrinsic interest to children and for their educational value. From a computerized record of a student's use of AD-1 it should be possible to assess ability in spelling, vocabulary size, or knowledge of multiple meanings and, if desired, reduce those assessments to comparisons with other students or with external criteria.

A third line of research would concentrate on the development of more sophisticated models. Two alternatives suggest themselves a priori. One line would explore programs for using the dictionary data base in a variety of different ways of presenting tasks and keeping records. A second line would explore additions to the data base of various kinds.

As an example of the first kind, it might prove desirable to be able to key the dictionary to particular texts that children were reading. If a new word appeared in the text, the teacher might wish to display only that sense of the word that was appropriate in the context, thus saving children the difficult task of selecting the appropriate sense. Special software would be required to enable the teacher to make such modifications in the standard display. Or it might be desirable to edit the dictionary in other ways -- simplifying definitions, inserting more pictures, giving more sentences illustrating usage, deleting etymology, or whatever -- which would also require special software. Or the various tasks (exemplified above) might be presented by the system, rather than the teacher, which would also complicate the system.

Such revisions, leading to what we might call AD-2, should be undertaken in close collaboration between systems programmers and educational research workers familiar with the potential uses of the system. It is not impossible to imagine elaborate and expensive software innovations that would have no real value to teachers or students.

Versions that might be called AD-3 should consider additions to the data base itself. Perhaps the most obvious example would be to add a thesaurus to the dictionary, thus providing cross-references between lexical entries that are valuable in searching for apposite words. Alphabetic versions of Roget's thesaurus are probably adequate to support initial explorations in this direction at the present time.

A request for information from a thesaurus elicits a set of words. Computers can do interesting things with sets that would be quite laborious to do by hand. They can, for example, find the intersection of two sets. Thus, it should be possible to type two words into AD-3 and have displayed only those words that are listed under both entries -- thus sharply narrowing the range of search.

Various teaching-testing tasks might be set on AD-3. For example, two words might be selected with nonintersecting sets of associates, and the student challenged to find the shortest path between them. A trace of the student's exploration should provide an interesting picture of his or her sensitivity to alternative meanings of words. Once the path was found, the student might be asked to return to the dictionary and select the particular senses of each word that are involved in the path. (It might also give the student an intuitive feeling for the theory of finite sets.)

Dictionaries do not itemize hyponyms. That is to say, under both 'table' and 'chair' they will give the superordinate term, 'article of furniture,' but if you look up 'furniture,' you will not find 'table' or 'chair.' (The exception is crossword dictionaries, which specialize in hyponyms.) This information could also be added to an automated dictionary (at least, for nouns). In some cases -- animal and plant taxonomies, for example -- this lexical information is part of the subject matter of some courses.

However, the most valuable additions to an automated dictionary -- grammar and general knowledge -- would also be the most difficult to make in the present state of the art. There is something artificial about any use of words in isolation. In order to deal with words as they occur in normal communication, the system must be able to recognize sentences and to assign interpretations to them that are reasonable, given the context and our presupposed general knowledge about the world and the events that occur in it. AD-4, incorporating grammar, and AD-5, incorporating grammar and general knowledge, are still beyond the reach of current science and technology. Enough progress has been made in these directions, however, that we can at least speculate about such systems and their educational implications.

It would be possible to add that component of a grammar known as the "phrase structure rules" directly to the entries in an automated dictionary, simply by adding to each entry the grammatical contexts in which that word can occur. Such a representation would be enormously redundant (the rule for forming noun phrases, for example, would be repeated in the entry for every noun), and there are reasons to believe it could not be complete (the transformational component of the grammar cannot be captured lexically). The only scheme that is considered plausible at the present time is to add grammatical rules as a component separate from the lexical component, in which case the system should probably be referred to as an automated grammar, rather than an automated dictionary.

Relevant evidence might be obtained, however, by a careful analysis of children's errors. If those errors can be characterized in terms of syntactic classes of words, rather than in terms of individual words, it would provide strong evidence for an independent grammatical component and against the incorporation of even phrase structure rules in the lexical component. Evidence available at the present time suggests that children learn grammatical rules abstractly, not redundantly for each particular word, because they often overgeneralize a new rule.

An automated grammar (AG-1) is feasible at the present time, given the various parsing systems that have been developed by computational linguists. These systems have frequently been used in question answering systems: the question is parsed and compiled as a program for retrieving information from a prestored data base, then answered with simplified declarative syntax. None of these systems is capable presently of dealing with the full range of grammatical English sentences, and most can process only a small vocabulary of terms directly related to the data base and the questions that can be asked of it.

AG-1 could support two lines of research, one leading to more sophisticated systems, the other exploring its educational applications. Along the first line, it would be useful to see whether a dictionary could provide an interesting data base for a question answering system. Along the second line, one would attempt to devise educational tasks using AG-1:

-- Parsing. The student tries to parse a simple sentence (one known to be correctly analyzed by AG-1) by typing its labelled bracketing. If the student's parsing is correct, it is accepted; if not, various responses might occur -- rejection, acceptance of only the correct parts, suggestions as to what is wrong, etc.

-- Syntactic ambiguity. Some sentences can be parsed in more than one way. Students might compete to find all correct parses, with AG-1 serving as referee.

-- Sentence anagrams. The student could be challenged to form as many sentences as possible from a given set of words, and AG-1 used to settle whether any disputed ordering was grammatical.

-- Privileges of occurrence. The student could be given a sentence with one word missing and asked what words could fill the blank; AG-1 would accept or reject, relative to that sentence, any word that the student typed. The task should teach a distinction between sentences that are syntactically ill-formed and sentences that are grammatical, but semantically anomalous.

-- Editing. AG-1 should be able to do most of the relatively mechanical corrections that teachers make in student compositions. The advantage, in addition to freeing teacher time for instruction that AG-1 could not provide, would be immediate feedback to the student while his or her communicative intentions were still operative.

-- Question answering. Questions about words would be phrased in standard English, so that the various tasks developed for the automated dictionary would not have to be posed in some special code.

-- Future developments of AG-1 might lead to a linguistic equivalent of the arithmetic BUGGY system, but we are presently far from understanding children's mistakes well enough to simulate their alternative grammars in a computer.

We are even further from AD-5, a system that could interpret sentences in the light of general knowledge about the world. As a metaphor, it is easy to speak of adding an encyclopedia to the dictionary in the system's data base, in which case the system should probably be called an Automated Encyclopedia (AE-1). Not until the representation of large bodies of information in a computer memory is better understood, however, will it be feasible to implement such a system. Preliminary research might begin, however, with particular bodies of information relevant to specific academic subjects. Question answering systems with a data base on, say, the American Revolution, might be experimentally feasible at the present time.

A full scale Automated Encyclopedia could be a powerful teacher, and might be taken as a distant goal to work toward. Meanwhile, systems at the level of AD-1 to AD-3 could not only support interesting educational research, but could provide valuable experiences to guide the development of more sophisticated systems.

III. TESTING

SOME EMERGING TRENDS IN TESTING

Norman Frederiksen
Director, Division of Psychological Studies
Educational Testing Service

I have chosen five topics: (1) tailored testing and other applications of latent trait theory, (2) alternatives to multiple-choice tests, (3) some new approaches to test validation, (4) contribution of tests to the understanding of cognitive processes, and (5) the use of tests in teaching. I'm not sure that all five really represent emerging trends in testing, but they at least represent ideas that I hope will influence testing in the years to come.

Tailored Testing and Other Applications of Latent Trait Theory

Fred Lord showed a long time ago that a conventional test cannot measure accurately the abilities of examinees at both the high and low ends of the score distribution for a heterogeneous group. Difficult items are needed to measure ability at the high end of the score distribution and easy items at the low end of the distribution. Hard items not only waste the time of the low-scoring candidates but also impair the measurement based on appropriate items. In order to measure accurately throughout the range of ability, it is necessary to match the difficulty level of the items to the ability level of each examinee. Tailored testing makes such individualized testing possible (Lord, 1977).

Vern Urry (1977) describes tailored testing as follows:

> The examinee is seated at a cathode ray tube terminal. A test item appears on the...screen..., and the examinee indicates his answer on a typewriter-like keyboard. If the answer given is correct, the next item will be more difficult; if the answer given is not correct, an easier item will follow. With each response, the computer makes a revised estimate of the examinee's ability; each revised estimate becomes more reliable. The test is terminated when the estimate reaches a specified level of reliability.

He goes on to say:

> Under certain conditions, tailored testing leads to dramatic improvements in measurement, compared to conventional paper-and-pencil testing. If one ability bank [of test items] is available: (a) the degree of reliability can be controlled; and (b) the number of items required for reliable measurement can be substantially reduced. If several ability banks [of items, for several tests,] and an external criterion are available: (a) the multiple correlation between the several tailored tests and an external criterion measure can be main-

> tained at an asymptotically high level; and (b) the number of items required for valid multidimensional measurement can be substantially reduced.

Urry (1977) has been developing tailored testing procedures for use by the Civil Service Commission. He reports that the cost of tailored testing is expected to be less than that of conventional testing because of the reduction in number of items needed for valid measurement. The remaining tasks involved in making the system operational for the Civil Service Commission include acquisition and deployment of hardware, training examiners, and changing regulations regarding examining practices.

How is it possible to give different forms of a test, all of different levels of difficulty, to subgroups of a population of examinees and obtain comparable scores, as though all candidates had taken the same test? Tailored testing is possible because of the development of latent trait models, most notably by Fred Lord (1952), whose monograph, "A Theory of Test Scores," was published more than 25 years ago. Central to latent trait models is the idea of an item characteristic curve, which is a mathematical function that relates the probability of success on an item to an unobserved latent trait or ability measured by the test as a whole. The parameters of an item characteristic curve represent item difficulty, item discrimination, and probability of guessing correctly. Since the shape of an item characteristic curve is invariant across different nonrandom samples of an examinee population, the model provides a way to compare examinees even though they have taken different subsets of test items. Two examinees who receive identical raw scores on easy and hard tests will nevertheless differ in the estimates of their ability on the latent trait. By fitting test data to the latent trait model, solutions can be found to testing problems not solvable by classical test theory.

Tailored testing is just one application of latent trait theory. Whether or not computerized tailored testing becomes economically feasible, the theory makes possible many improvements in the testing field. The entire 1977 summer issue of the Journal of Educational Measurement is devoted to applications of latent trait models, which are described by Gary Marco (1977) as "a quiet revolution in the field of educational measurement." The editor of JEM comments that the domain is virtually unknown to measurement practitioners, although most problems of tractability have been solved and the usefulness of the models in handling educational measurement problems cannot be disputed. The issue of JEM is intended to help solve the problem of diffusion of information. It provides an excellent introduction to the concepts and assumptions of latent trait theory and many examples of applications, including use of latent trait models in selecting items for a conventional test, redesigning a test for specific purposes, criterion-referenced testing, mastery testing, two-stage testing (an alternative to tailored testing), differential weighting of response alternatives in multiple-choice test items, identifying individuals among a group of examinees for whom the test is not appropriate, studying item bias, and equating tests on the basis of item statistics obtained from different groups.

A book on applications of item characteristic curve theory by Fred Lord is in preparation. The publisher, Lawrence Erlbaum Associates, expects that it will be available by mid-1979. Computer programs are available. The application of latent trait models is clearly an emerging trend in the field of testing that will make possible important advances in efficiency and accuracy of measurement.

Alternatives to Multiple-Choice Tests

Latent trait theory is useful for tests containing items that can be scored as right or wrong and that measure a unitary trait or ability. Such tests are most likely to be of the multiple-choice form. Many criticisms of testing center on the use of the multiple-choice format. The usual criticism is that the examinee doesn't have to remember anything, he must only be able to recognize the answer. The criticism, at least in that form, is not necessarily justified, since it is possible to write multiple-choice items that call upon complex processes of inference and reasoning. Multiple-choice tests can be very efficient and useful in assessing not only a student's knowledge and understanding, but also his ability to solve problems involving algorithms and reasoning from facts to conclusion. I do not think we should abandon machine-scorable formats such as the multiple-choice.

But real-life problems do not ordinarily appear in multiple-choice form with all the options clearly presented. More likely, a problem will appear, initially, merely as a feeling of dissatisfaction or discomfort about something, and it may be necessary to discover more accurately what the true nature of the problem is before it can be solved. Once the problem has been more clearly formulated, solving the problem may require a series of steps such as (1) generating ideas (call them hypotheses) about possible causes or solutions; (2) deciding which of the ideas are worth exploring; (3) thinking of sources of information that might suggest additional hypotheses or help you decide which hypotheses should be dropped and which investigated; (4) getting relevant information and considering what inferences may be drawn from it that would help in testing the hypotheses; (5) eventually choosing an attractive hypothesis; (6) devising ways of ascertaining whether it is correct or not by mental or "experimental" operations; (7) carrying out such operations; and (8) going through additional loops in this process until a satisfactory solution is found, or you give up. Multiple-choice tests can no doubt measure much of the knowledge and some of the skills involved in the process, but they will certainly not reflect the whole problem-solving procedure. I believe we should develop testing procedures that will assess aspects of the thinking process not adequately dealt with by multiple-choice tests.

Suppose we were somehow able to erase from the minds of testers all their preconceptions about tests; how would they be likely to go about evaluating the performance of a student? My guess is that they would do more sensible things than write multiple-choice items. To measure what a student had learned about spelling, they would probably give him words to spell; and to investigate how well he had learned to criticize scientific papers, they might give him a scientific paper to review. My definition of a test follows this notion: a test is any standardized procedure for eliciting the kind of behavior one wants to observe--and I mean the kind of behavior one really wants to observe. This definition frees us from any preconceptions regarding test formats or item types, particularly the preconception that a test item requires a presentation of five possible answers. The definition leads one to devise whatever situations he believes will best lead the candidate to exhibit the performance one wants to evaluate. The definition may also lead to a variety of methods for assessing the performance observed. It may be possible, instead of merely counting the number of correct answers, to assess quality of responses from each of several points of view, or to make judgments about strategies or styles of responding. There may be many important dimensions of performance other than the number of correct answers.

The enterprise of constructing tests that do not fit the usual conception of a test is not the free-wheeling activity that may be implied by the definition. The need for tests that are sound psychometrically is no less than for multiple-choice tests. But a certain amount of ingenuity and willingness to try new paths is necessary for one who embarks on more adventurous kinds of test development.

I first became sensitized to the potential problems of multiple-choice testing during World War II when I was a member of a group, led by Harold Gulliksen, that worked on personnel selection and training in the Navy. I was sent to Bainbridge, Maryland, to try to improve the evaluation of instruction in a Class A Gunner's Mate school, a school where students were taught to maintain and repair the naval guns aboard a warship. We found that Gunner's Mate school grades were based on multiple-choice tests dealing largely with names and functions of gun parts, such as the breech block locking bolt. When we suggested that performance tests would be more appropriate for evaluating ability to analyze casualties and repair guns, we were told that it was impossible because it would take too long, and besides, the students couldn't do it. However, we had orders from the Bureau of Naval Personnel, and it was therefore possible for us to develop tests that required students to perform such tasks as adjusting the oil buffer of a .50 caliber Browning machine gun, removing and replacing the interlock carrier spring from a 20 MM gun, and removing and replacing the extractor plunger on a 5"/38 antiaircraft gun. The instructors were right; the students couldn't do the tasks. But the introduction of the tests resulted in some fundamental changes in the school. I'll refer to these changes later.

Another example, this one of a more academic sort: The board that oversees the development and use of the Graduate Record Examinations was interested a few years ago in the possibility of developing tests that might measure that intangible characteristic "creativity" that is assumed to be of great importance in the academic world, whether it involves art, literature, or science. Bill Ward and I accepted the challenge and offered to try to devise methods for measuring creativity in a limited sense--the creativity that might be involved in scientific problem solving. In order to reduce the odds against us, we decided to work in the area of behavioral science, which we know best.

After considering a number of possibilities, we came up with four free-response paper-and-pencil tests. Their names are Formulating Hypotheses, Evaluating Proposals, Solving Methodological Problems, and Measuring Constructs. The names give an indication of the nature of the tests, but let me describe one of them in a little more detail.

Each Formulating Hypotheses problem is a brief description of a scientific investigation. The results of the investigation are shown in the form of a graph or table, and the major finding is clearly stated--for example, "After viewing television programs with aggressive content, the children showed increases in interpersonal aggression." The examinee's task is to suggest hypotheses that might explain the finding--not only the hypothesis he thinks is most likely to be correct, but also others that should be considered in interpreting the data or planning another investigation. There is no list of hypotheses--the candidate must think of them and write them down. He later marks the idea that he thinks is the best hypothesis.

The most obvious method of scoring such a test is to hire a squadron of research psychologists to sit down and read the responses and assign them grades. We used a somewhat different method that we think has a number of advantages over the essay-grading method. We call it category scoring. The idea is, first, to make a classification of the ideas actually written by examinees and to write a definition of each category. Then the scorer's task is merely to assign each response written by an examinee to one of the categories. Members of a panel of expert judges are asked to make judgments of the quality of each category, and scale values are derived. The scale values can be stored in a computer, and when the category numbers for each examinee are fed to the computer, it can generate a variety of scores. We have worked with scores that reflect the average quality of all the examinee's ideas, the quality of the ideas he thinks are best, the quality of ideas that are best by our scoring standards, the number of ideas, the number of unusual ideas, and the number of ideas that are both unusual and of high quality. The latter might be considered a kind of creativity score.

A tryout of Formulating Hypotheses and the other three tests showed that in most instances a test based on six problems is reasonably reliable and is suitable in difficulty for candidates for admission to graduate school or for first-year graduate students. Scores based on quality of responses are quite independent of scores based on number of responses; and the scores, especially the number scores, have very low correlations with GRE tests, including Verbal and Quantitative ability scores and the Advanced Psychology Test score. A follow-up study of those candidates who actually attended graduate school in psychology showed that the scores based on number were potentially better predictors of professional accomplishments in the first year of graduate school than the GRE tests. The accomplishments were such early indicators of professional performance as number of publications (as author or co-author), amount of participation in collaborative research, and number of professional activities reported.

I'll give additional examples of nonmultiple-choice tests later. I hope I have communicated the idea that it is possible to present problems in formats that require more than merely comprehending and evaluating options presented, and that reflect other abilities such as those involved in creating and evaluating hypotheses.

Approaches to Test Validation

Start with Criterion Development

The usual practice in test construction is to begin by developing a test and then to look for some available criterion variable for use in computing a "validity coefficient." Thus we have a host of tests used in educational assessment and selection that are validated by computing correlations between scores on the test and first-year grades--college grades, law school grades, medical school grades, or whatever. Most of the educators who use the tests in selecting students undoubtedly hope or believe that the tests have something to do with success after the training is completed, but the test publishers are usually careful to avoid making claims beyond the prediction of first-year grades. Sometimes a daring researcher makes a study of "long-term" predictive validity by computing correlations with grades in senior year; but this practice is not very reinforcing because the correlations with senior grades usually turn out to be low.

Studies in medical schools showed quite clearly that the Medical College Admissions Test (MCAT) predicted grades in the first two years of training pretty well but had very little predictive value for the third and fourth years. These are the clinical years when students are working with patients to acquire skills in medical diagnosis and patient management. Evaluation of performance in the clinical years is typically based on judgments made by professors who observe the behavior of the students who accompany them on their hospital rounds. Each student may be asked to read the chart of a patient, do parts of a physical examination, talk to the patient, and make comments about diagnosis and treatment. Although observations tend to be unsystematic and professors may not agree with each other very well, the cumulative judgments undoubtedly have substantial reliability. The low test validity for the clinical years presumably reflects the inappropriateness of the selection tests for predicting clinical performance.

A completely different approach to the test development procedure is possible. Instead of beginning with selection tests, why not reverse the order and begin with development of criterion measures? There are two advantages of using such a procedure: one is that when the selection test is completed, criterion variables will be available for concurrent or predictive validation studies. The second advantage is more important: the careful development of criterion measures should provide a much better understanding of the criterion domain and thus a better basis for deciding what the selection tests should be like.

Criterion development should begin with no preconceptions about test format or testing procedures. The goal should be to devise standardized procedures for eliciting the behaviors one wants to observe and measure. In many cases, the best procedure might be a test based on realistic simulations of problem situations that are typical of those the student is being trained to deal with in real life.

The first step in criterion development is obviously to study the criterion domain. The procedures will vary depending upon the nature of the domain. If it is an area of educational achievement such as American History, one might want to make a survey of representative textbooks, syllabi, and other instructional materials, supplemented by judgments made by a representative sample of educators in the field regarding the relative importance of the elements found to make up the domain. A more interesting problem involves tests intended to select candidates for admission to professional schools such as law and medicine, where the purpose is to select those who are likely to succeed in the sense of being good lawyers or doctors as well as to earn high grades in professional school courses. Here the study of the criterion domain would have to be some sort of task analysis, or perhaps a Flanagan (1954) critical incidents study, that identifies the most critical tasks to be performed. If the procedures used in studying the criterion domain and translating the results into a test are well conceived and carefuly executed, we have some confidence that the test to be developed will represent the domain.

The next task is to devise procedures for eliciting behaviors that can be evaluated in terms of these critical elements of job performance. The procedure I suggest is to make lifelike simulations of problem situations.

A project of this sort is now under way. It is concerned with improving the selection of medical students. The project involves the collaboration of the National Board of Medical Examiners (NBME), ETS, and a consortium of medical schools, and is funded by the Robert Wood Johnson Foundation. One specific purpose is to select students who will eventually do well in diagnostic problem solving. The plan of the study is to begin with criterion development, and, on the basis of experience in that area, to develop experimental selection tests.

The criterion tests (developed by NBME) are paper-pencil simulations of diagnostic problems. The first page of a test item briefly describes a situation that might present itself to, say, a resident on duty in the emergency ward of a hospital. A patient is brought in complaining of chest pains. The initial report shows that he is a 55 year old male, a laborer, and a heavy smoker, and that the pains began during the afternoon. On the basis of this scanty information, the subject is asked to write his first impressions regarding possible diagnoses--heart attack, pneumonia, pleurisy, a strained muscle, etc. On the following page, the examinee is asked to indicate what he wishes to do next. He may want to interview the patient, do a physical examination, order laboratory tests, or even initiate treatment. If the examinee says he wishes to order laboratory tests, he is directed to turn to the appropriate page of the test booklet, where he orders the tests he wants and receives the results of the tests by using a felt pen that makes a latent image visible. Then the examinee is given an opportunity to write the diagnostic hypotheses he now thinks should be considered. Then he chooses his next course of action. And so the test proceeds until the examinee is ready to write the diagnosis and to prescribe the initial steps in his treatment plan.

Bill Ward and I have developed at ETS a set of problem-solving tests that are prototypes of possible selection tests. They are intended to parallel the medical tests except that they do not require medical knowledge and experience. Like the medical tests, they are unfolding problems.

The examinee begins by reading a paragraph that describes his assumed role (e.g., a new assistant to the Health Commissioner of Iron City) and presents information (age distributions) from which a problem can be inferred (why is there a smaller proportion of elderly people in Iron City than in the state as a whole?). On the basis of a small amount of information, the examinee is asked to write whatever hypotheses come to mind that might explain the smaller proportion of old people (diseases caused by air or water pollution, poor sanitation, inadequate health services, occupational disease, migration of old people out of the city, migration of younger people into the city, etc.). On the next page he is told that the Department of Health has information that may be relevant to his hypotheses, and he is asked to write the items of information he will ask for. He is also told that the Bureau of Vital Statistics has a tape that makes it possible to make breakdowns of the Iron City age distributions, and he is asked what breakdowns of the data he will request. Information resulting from these requests is presented, and the subject is given an opportunity to present a new list of hypotheses--and so it goes until the subject is asked to indicate what he is going to report to the Commissioner.

The unfolding problems offer opportunities for generating not only scores like those we obtain from Formulating Hypotheses, but also a number of additional scores that describe sequences in problem solving behavior. Scores can be obtained for each page separately (and later combined in various ways) that might, for example, contrast performance in early and late stages in the process of solving the problem. We can count the number of ideas that are justifiably dropped after new information is available--the number of "good" drops. We can also count the number of "bad" drops and the number of "good" and "bad" retentions of an idea. We can tabulate frequencies of use of various presumed sources of ideas--whether the ideas are directly suggested by items of information received, whether they are apparently based on inferences involving the information, or whether they are neither--"creative" ideas that have no obvious origin in information provided. It is also possible to obtain scores that reflect the breadth of the candidate's thinking by scoring for the number of category groups he uses (e.g., pollution, health services, migration, etc.) as compared with, for example, listing six possible sources of pollution.

I have tried to show that the usual practice of starting with test development and seeking easily available criteria for use in validation studies is likely to lead to the development of inappropriate tests, since the criterion variables that are easily available are unlikely to reflect very accurately the criterion domain; and that it is feasible to set out initially to develop criterion measures that more accurately represent the constructs one wishes to measure. Criterion development should be based on an appropriate study of the criterion domain and might in many instances consist in making simulations of the essential aspect of situations of the sort students are being trained to deal with. Undertaking criterion development before developing tests should result in tests that are more likely to be valid, whether they are used in assessing school achievement, in guidance, or for selection.

Systematic Construct Validation

The fact that one has developed something called a criterion measure, of course, does not mean that everyone will agree that it is a good criterion. It is certain that someone will challenge the claim. What do you do? Create a still more elaborate and realistic ultimate criterion measure against which to validate the first one? Maybe, but it is obvious that this could lead to an infinite regress of criterion building. A solution other than criterion-related validity is obviously needed. Construct validity provides an answer.

Construct validity is defined by Messick (1975) as "... the process of marshaling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning." Evidence for construct validity would show that the measure in question relates positively to those variables it should relate to on theoretical grounds (convergent evidence) and that the measure is not unduly related to other distinct constructs (discriminant evidence).

How does one go about establishing construct validity? The writers of the Standards for Educational and Psychological Tests (APA, 1966) say that "Evidence of construct validity is not found in a single study; rather, judgments of construct validity are based upon an accumulation of research results." It is, of course, true that if enough investigators happen to be interested enough in a particular instrument to examine its relationships to other measures, to administer it to contrasting groups, to vary certain aspects of the test or the conditions under which it is given, relevant information will accumulate. But the information will be unsystematic, since each researcher explores his own hypotheses; and furthermore, the data may be slow in accumulating or may not accumulate at all. The statement in the Standards implies too little responsibility on the part of the test developer for providing evidence of construct validity. I suggest that construct validity studies should be carried out systematically by the test developer. I would like to illustrate such a systematic study by describing our plans for studying the construct validity of the medical school selection and criterion tests.

The basic plan was as follows: Two kinds of measures were to be developed--criterion measures and experimental selection measures. The criterion measures are the paper-pencil simulations of diagnostic problems that I have described. The selection tests are problem-solving tasks based on simulations of nonmedical situations that are supposed to parallel the medical problems except that they do not require knowledge and experience in medicine.

A third set of tests was also to be included. This is a set of cognitive and personality variables chosen for use in construct validation because of their theoretical relevance to the criterion and selection measures. We call them process variables. They include three tests to measure each of eight cognitive factors (a verbal factor, several reasoning factors, a cognitive-flexibility factor, and two "divergent production" factors) and an inventory that yields scores reflecting personality characteristics that might be useful in helping to account for variations in behavior. Also included are scores on conventional tests, grades in college or medical school, and other items of information that are frequently used by admissions committees. Finally, scores on NBME tests of medical knowledge will be available.

The criterion measures, selection tests, and the process-variable tests were all administered to fourth-year medical students. The selection tests and process variable tests, but not the criterion tests, were given to first-year medical students. The criterion tests were given to third-year residents. The data make possible several approaches to our problem of validating the new tests:

1. If we assume that the criterion variables are valid, we can obtain evidence of concurrent validity of the selection tests by computing correlations between the criterion measures and the selection tests, using data for fourth-year students.

2. Again assuming validity of the criterion tests, we will, in the future, be able to obtain evidence of predictive validity by a longitudinal study of the first-year students.

3. Using fourth-year student data, we can obtain information relevant to construct validity of the criterion measures by examining the relationships between the criterion scores and the process variables. We hypothesize that performance on the diagnostic problem-solving tasks requires not only knowledge of medicine but also such abilities as verbal comprehension, inductive and logical reasoning, cognitive flexibility, and (especially in the early phases of problem solving) ideational fluency. Expressional fluency will also be required to some degree, since responses must be composed. Some of the scores might be related to self-confidence, anxiety or defensiveness, acceptance of complexity, tolerance for ambiguity, and the like. To the extent that the correlations are consistent with reasonable hypotheses, our confidence in the new measures of diagnostic ability may increase--in other words, construct validity is to some degree established.

4. In exactly the same way, using the same process variables (except tests of medical knowledge), we can look for relationships supporting the construct validity of the selection tests. This can be done separately for first- and fourth-year students, which makes possible a comparison of the constructs for students who differ in amount of medical training.

5. Since our theoretical expectations are the same for criterion variables and selection test scores (except that knowledge of medicine is presumably not a factor in selection test performance), we can (using fourth-year student data) see how similar the selection and criterion variables are with respect to their relationships to the process measures. If the patterns of relationships are very similar, we can conclude that selection tests and criterion tests measure the same construct. This is a kind of validity that hasn't been named, so far as I know. Call it a "construct similarity."

6. Similarly, it will be possible to investigate the construct similarity of the experimental tests to such conventional tests as the MCAT, the NBME tests, and to other items of information used in admissions decisions, such as undergraduate grade-point average.

7. Comparisons of score distributions obtained from fourth-year students and third-year residents will reveal whether scores tend to change in directions that would be predicted on the basis of three years of experience in the practice of medicine.

I have tried to demonstrate by this example that it is possible to design investigations that will provide a great deal of information relevant to judging construct validity, and that one need not wait for an accumulation of research results in the journals.

Use of Tests in Investigating Cognitive Process

As Sam Messick (1975) said in his Division 5 presidential address, "... test validation in the construct framework is integrated with hypothesis testing and with all of the philosophical and empirical means by which scientific theories are evaluated." In other words, a study of the construct validity of a test may be equivalent to a scientific investigation of the cognitive process involved in taking the test. Some psychologists who use information-processing concepts and methodologies have in fact become interested in studying the cognitive processes that are involved in taking test items of the kind typically found in psychological tests.

Since the operations involved in construct validation may include experimental manipulations as well as correlational methods, experimental psychology and correlational psychology may profitably join hands in this enterprise (Cronbach, 1955). A paper by MacLeod, Hunt, and Mathews (1978) provides an interesting example of how experimental and psychometric approaches can be combined. They used the Chase and Clark (1972) sentence-picture comprehension task. The subject is seated before a display scope where he sees a sentence (e.g., PLUS IS ABOVE STAR) and then a picture, which is either a star above a plus sign or a plus sign above a star. The task is to indicate as quickly as possible by pressing a button whether or not the sentence is true with respect to the picture. The dependent variable is reaction time. In the MacLeod, Hunt, and Mathews study, two components of the process were separately timed. The first component was the time required to read and comprehend the sentence, and the second was verification time--the time required to decide whether the sentence matched the picture.

On the basis of differences between these two latency measures, two groups of subjects were identified. Members of one group had short reaction times for comprehending the sentence and relatively long reaction times for verification; the second group was characterized by longer reaction time for comprehension and shorter time for verification.

The first group was presumed to have used a linguistic strategy. In the comprehension period they formed an internal representation of the sentence; and in the verification period, they (1) formed an internal representation of the picture, (2) converted it into a sentence representation, and (3) compared the two sentences. The second group was thought to have used a pictorial strategy. In the comprehension period, they (1) formed an internal representation of the sentence and (2) converted it into a representation of a picture. In the verification period, they (1) formed a representation of a picture and (2) compared the two pictures. It was found that subjects using the second method, in which pictures were compared, were markedly higher in spatial ability than those who compared sentences. Verification time correlated with scores on a verbal ability test for the first group and with scores on a spatial ability test for the second group. Thus the psychometric data were used in verifying an interpretation of an experimental result. The results show that people with different cognitive abilities used quite different processing strategies in solving the problems.

Another example illustrates how information generated for the purpose of construct validation may be useful in studying process. The purpose was to compare two forms of Formulating Hypotheses with respect to construct validity. In one form the items were in free-response format, the other used a sort of multiple-choice format that is scorable by machine. Free-response vs. machine-scorable format may be thought of as an experimental treatment. The machine-scorable version contains lists of hypotheses instead of the blank spaces that occur in the free-response form, and the subjects' tasks are (1) to mark all the hypotheses that ought to be considered, and (2) rank all the hypotheses in order of quality. Analogous scores were developed for the two forms, except that a seventh score, rho, was computed for each machine-scorable problem; it is the rank correlation between the subject's rank order and the ranking based on the judgments of our panel of experts.

Process variables much like those I described for the medical school study were administered along with free-response and machine-scorable FH items. Subjects were college seniors who planned to apply for admission to graduate school in psychology. Separate analyses were performed for cognitive and for personality process variables. In both instances, a Jöreskog maximum likelihood factor analysis was used to test the fit of the factor model to the data. Good fits were obtained for both analyses.

Scores on the two forms of Formulating Hypotheses were treated as extension variables, so that their loadings on all the factors could be obtained without their influencing the factor structure.

Table 1 shows the results for three personality factors. Extension loadings are below the double line. If we compare extension loadings for scores on the free-response and machine-scorable forms of Formulating Hypotheses, we see that the major difference is that number scores, for the free-response form only, have loadings on the Maladjustment factor. Subjects who are anxious or defensive perhaps write more hypotheses in order to avoid looking stupid or to make a good impression. Or anxiety may operate as a drive (Taylor, 1956). As the loadings indicate, the anxious people also tend to write not only more ideas but also more ideas that are both unusual and of high quality.

Table 1

Maximum Likelihood Factor Analysis of SoAnDe with Extension Loadings

		Social Extraversion I	Maladjustment II	Acceptance of Complexity III
Dominance		.83	0	0
Impulsivity		.55	0	0
Sociability		.68	0	0
Anxiety		0	.79	0
Maladjustment		0	.77	0
Stability		0	—.75	0
Preference for Complexity		0	0	.79
Rigidity		0	0	—.50
Tolerance for Ambiguity		0	0	.51
FR	Best Quality	—.19	.10	.23
	Mean Quality	—.18	.09	.24
	Highest Quality	—.21	.15	.24
	Number	—.11	.30	.14
	Unusual	—.07	.32	.00
	Unusual-High	—.09	.35	—.02
MS	Best Quality	.00	.00	.22
	Mean Quality	.01	—.02	.05
	Highest Quality	—.04	.04	.23
	Number	—.07	.12	.18
	Unusual	—.05	.10	.17
	Unusual-High	—.03	.07	.17
	Rho	—.07	—.02	.20

Table 2 shows the results for cognitive factors. The major finding, which was of course predicted, is that the divergent production factors (Expressional Fluency and Ideational Fluency) are related to the three scores based on number of ideas for the free-response but not the machine-scorable form.

Table 2

Maximum Likelihood Factor Analysis of Cognitive Variables with Extension Loadings

	Vocab. I	Quantitative Reasoning II	Induction III	Logical Reasoning IV	Cognitive Flexibility V	Expressional Fluency VI	Ideational Fluency VII	Knowledge of Psychol. VIII
Vocabulary	.93	0	0	0	0	0	0	0
Word Meanings	.88	0	0	0	0	0	0	0
Math. Aptitude	0	.78	0	0	0	0	0	0
Necess. Arith. Opers.	0	.72	0	0	0	0	0	0
Resourceful Arith.	0	.85	0	0	0	0	0	0
Letter Sets	0	0	.71	0	0	0	0	0
Locations	0	0	.66	0	0	0	0	0
Figure Classif.	0	0	.68	0	0	0	0	0
Logical Diagrams	0	0	0	.65	0	0	0	0
Nonsense Syllogs.	0	0	0	.54	0	0	0	0
Inferences	0	0	0	.62	0	0	0	0
Reversed Reading	0	0	0	0	.84	0	0	0
Sign Changes	0	0	0	0	.64	0	0	0
Making Sentences	0	0	0	0	0	.54	0	0
Arranging Words	0	0	0	0	0	.69	0	0
Rewriting	0	0	0	0	0	.38	0	0
Pattern Interp. 1	0	0	0	0	0	0	.62	0
Diff. Uses	0	0	0	0	0	0	.77	0
Unexpected Results	0	0	0	0	0	0	.70	0
Adv. Psychol. 1	0	0	0	0	0	0	0	.80
Adv. Psychol. 2	0	0	0	0	0	0	0	.90
FR Best Quality	.23	.24	.30	.31	.26	.14	.04	.34
Mean Quality	.25	.25	.38	.41	.23	.17	.04	.45
Highest Quality	.20	.31	.44	.41	.26	.31	.18	.47
Number	.20	.31	.36	.39	.26	.44	.40	.36
Unusual	.10	.22	.26	.23	.20	.38	.36	.21
Unusual-High	.13	.23	.25	.18	.16	.24	.16	.28
MS Best Quality	.21	.20	.37	.30	.33	.18	.07	.47
Mean Quality	.12	.18	.18	.25	22	.19	.04	.06
Highest Quality	.16	.29	.47	.46	.32	.29	.15	.47
Number	.08	.16	.29	.22	.09	.07	.06	.39
Unusual	.04	.06	.16	.11	—.01	.01	.07	.34
Unusual High	.14	.25	.30	.29	.14	.08	.07	.46
Rho	.22	.34	.40	.49	.32	.30	.12	.35

A related finding is that Factor V, Cognitive Flexibility, relates to almost all the scores except number scores for the machine-scorable form. The tests marking this factor were devised by Scheier and Ferguson (1952; J. Frederiksen, 1967). They are supposed to measure ability to break sets. One, Reversed Reading, requires the examinee to read sentences in which the letters in each word are printed in reverse order; and the other, Sign Changes, requires him to solve simple arithmetic problems under instructions to add when there is a minus sign, multiply when there is a division sign, and so on. These tasks certainly do not involve reasoning. Yet they do relate to performance on the complex tasks required by Formulating Hypotheses items. Furthermore, the Cognitive Flexibility factor correlates highly with all the reasoning factors. Changing set--getting rid of ideas as well as getting ideas--seems to be involved in the problem-solving tasks, perhaps because of its importance in reasoning more generally.

I have tried to show by these examples that construct validation can lead to new scientific insights as well as contribute to decisions about how tests are to be used. Doing applied research in ways that contribute to understanding may result in better applied research as well as in contributions to the scientific study of a phenomenon.

Use of Tests in Teaching

There seems to be little doubt that tests influence the behavior of people in the classroom, both students and teachers. The first question you are likely to get from students on opening day is, "Will the exams be multiple-choice or essay?" The students want to know because their study methods depend on the kind of examination. The teacher is also likely to be concerned about the tests given to his or her students, perhaps because of fear that the results will make the teacher look bad. Even school administrators care. We had a superintendent of public instruction in my state a few years ago who favored banning all educational tests because they influence the behavior of teachers. I thought he was wrong. If I had been the superintendent and had in my grasp a tool I could use to influence the behavior of teachers, I would have held on for dear life.

I mentioned earlier the work I did in a Gunner's Mates school during World War II. We constructed a set of performance tests to supplement the multiple-choice tests. The new tests required students to disassemble guns, analyze casualties, make adjustments, and reassemble. As the Navy chiefs predicted, the students did very poorly. But because successive classes overlapped, new students soon got word of what the new tests were like, and they began practicing the disassembly and assembly of guns. Performance improved with each class. The instructors also got the point. They moved out the classroom chairs and lecture podium and brought in more guns and gun mounts. The tests soon became too easy. Incidentally, correlations with tests in the Navy's Basic Battery also changed. The validities for verbal and reading tests dropped, and mechanical aptitude and mechanical knowledge became the best tests for predicting grades in Gunner's Mate school.

No attempt was made to change the curriculum. A dramatic change in achievement came about solely through a change in the tests. It seems very clear that teaching and learning can be influenced by tests. It is also clear that those who make the tests have a great responsibility to produce tests that will influence teachers to teach and students to learn the knowledge and skills that truly reflect the objectives of the training program.

I said earlier that good multiple-choice tests can be made. This is not the same as saying that all multiple-choice tests are good. Most of such tests probably contain items that stress recognition of detailed items of information. Even the best multiple-choice tests by their very nature present options for choice rather than requiring examinees to think of the options for themselves. It seems to me that the heavy reliance on multiple-choice tests implies a substantial and important kind of bias that tends to exclude from education divergent thinking skills such as those required for clarification of problems, development of hypotheses, seeking appropriate information to test hypotheses, and judging when the process of seeking new hypotheses and new information may be terminated. Multiple-choice items by their very nature cannot elicit such behaviors to any significant extent. Introduction of tests that do elicit such performance would stimulate teachers to teach and students to practice such skills.

Assuming that it is a good idea to supplement the multiple-choice tests with tests that are more likely to elicit thinking skills of the sort I have described, how does one go about effecting such a change? So far as the large testing programs like those of the College Board, the law and medical schools, and the GRE are concerned, there would be tremendous difficulties. The testing organizations have set up highly efficient routines for processing tests that use machines capable of scoring answer sheets at the rate of 300 per minute. With the candidate volume and fee schedules involved in such testing programs, it is hard to see how tests that require reading and evaluating of protocols could ever be used on a large scale. Reactive computer systems have been developed but the difficulties in teaching a computer to understand natural language and respond appropriately are enormous.

The ultimate solution may be to put the tests into the classroom so that their cost can be justified primarily on the basis of training value rather than usefulness for assessment. The testing organizations could facilitate the change by turning to the development of scorable test-like materials that could be used as class exercises or homework and as the basis for class discussions. The scores would be useful primarily in providing feedback to students; but scores would also be available for all the conventional purposes of tests.

Consider how Formulating Hypotheses, for example, could be used in a graduate course in research methodology. An FH item could be assigned as homework. At the next class meeting, each student would be given the list of scoring categories and asked to record the number of the category that best matched each of his hypotheses. Then the scale values of the hypotheses would be revealed, and students could score their own performance on the exercise. Then a lively discussion might follow in which students could challenge the scale values and argue with each other about the value of their ideas, and the instructor would be prepared to provide information and arguments relevant to the evaluations. It would be possible also to provide in written form a discussion of the pros and cons of each category for use in situations where instructors could not be actively involved. Scores could be recorded and used as a basis for evaluating progress and assigning grades.

I would think that frequent use of such procedures would provide experiences in the more creative aspects of scientific thinking that graduate students tend not to get until they participate in advanced seminars with few students and the best research professors. Similar exercises could be prepared for use in high school and college, and probably also in elementary schools.

Summary

I have identified five emerging trends in educational measurement. The first is the use of latent trait theory, which now appears to be ready for use in improving many of the operations involved in test administration, scoring, equating, test interpretation, and so on, and which makes possible things that cannot be accomplished with classical test theory.

The theory, however, applies primarily to tests containing items that can be scored as right or wrong. Such tests are likely to stress knowledge and the kind of problem solving that requires thinking convergently from information to solution. I have proposed that such examination methods be supplemented by assessment methods that more adequately reflect divergent thinking skills, even though such tests cannot be scored with a machine.

I have suggested that validation of tests too often involves choosing criterion variables as an afterthought, and that one might better begin the test development process by developing criterion variables; and I recommended that we think of construct validation as a systematic effort to obtain information, using experimental and correlational procedures, to test hypotheses regarding the processes involved in taking a test. Construct validation can thus be viewed as contributing to the scientific study of cognition as well as to better testing.

Finally, I have suggested that since tests have a great deal of influence on teaching and learning, their use and cost can best be justified on the basis of their value in instruction.

Viewed from this standpoint, tests that elicit behaviors reflecting the full range of educational goals are indispensable components of the educational process.

References

American Psychological Association. Standards for educational and psychological tests and manuals. Washington, D.C.: Author, 1966.

Chase, W.G., & Clark, H.H. Mental operations in the comparison of sentences and pictures. In L. Gregg (Ed.), Cognition in learning and memory. New York: Wiley, 1972.

Cronbach, L.J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.

Flanagan, J.C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-348.

Frederiksen, J.R. Cognitive factors in the recognition of ambiguous auditory and visual stimuli. Journal of Personality and Social Psychology, 1967, 7 (1, Whole No. 639). (Monograph).

Lord, F.M. A theory of test scores. Psychometric Monograph, 1953, No. 7.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

MacLeod, C.M., Hunt E.B., & Mathews, N.N. Individual differences in the verification of sentence-picture relationships. Journal of Verbal Learning and Verbal Behavior, 1978, 17, 493-507.

Marco, G.L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Scheier, I.H., & Ferguson, G.A. Further factorial studies of tests of rigidity. Canadian Journal of Psychology, 1952, 6, 18-30.

Taylor, J.A. Drive theory and manifest anxiety. Psychological Bulletin, 1956, 53, 303-320.

Urry, V.W. Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

STANDARDIZED TESTING AND AFRICAN-AMERICANS: BUILDING ASSESSOR COMPETENCE IN SYSTEMATIC ASSESSMENT

Asa G. Hilliard, III
Dean, School of Education
San Francisco State University

In San Francisco, 5 African-American school boys were judged to be mentally retarded, mainly because of low scores on standardized I.Q. tests. They represent thousands of African-American children who are similarly mis-assessed. Allan Bakke is said to be "better qualified" for admission to medical school than minority applicants with lower scores, mainly because of his score on the Medical School Aptitude Test. Yet this is a test which has only a small positive correlation with first year medical school evaluations, and which has a low to negative correlation with assessments of later success in medical practice. In South Carolina, California, and other states, hundreds of African-American teachers, many of whom have been successful teachers, are being denied employment as teachers due to their scores on the National Teacher Examinations, even though the test is said not to be a measure of teaching performance. Hundreds of Head Start and day care programs are being evaluated for program and cost effectiveness based in large measure upon the scores which chidren get on such tests as the Peabody Picture Vocabulary Test, and the Pre-School Inventory, even though there is no systematic determination that these tests have content validity for the programs which are being assessed. I could go on. However, one thing is clear about standardized testing and African-American populations. They are, for the most part, the source of unconscionable gross errors in professional assessment which continue to have a negative effect on the lives of African-Americans.

Systematic assessment for African-Americans is as appropriate, as desirable, and as necessary as for any other group. However, mindless standardization in mass produced tests does not yet qualify as a scientific procedure. There is no requirement that assessment has to be standardized in order to be valid, any more than there is a requirement that language be standardized for communication to occur. The goal of any assessment practice must be to discover the truth and not to verify the rules of the assessment system. When the requirements of an assessment system actually generate error and falsehood as a common occurrence, then it is the system which must be assessed.

The potential for the improvement of standardized testing is tied to the possibility that those who produce standardized tests are able to wrench free from conceptual incarceration, the ideological prison which results in shallow consideration of basic issues. The most important constructs and assumptions which support the standardized testing movement have undergone no serious re-examination within the field since they were first posited. Psychologists, and educators in particular, have grown up with these constructs and assumptions to the point where, to some, the constructs and assumptions appear to be accepted as self evident. Consequently, discussions within the field of standardized testing by psychologists, educators and others tend to be very thorough--but simply on the wrong topics. The literature on standardized testing reveals increasing sophistication in the implementation of assessment practices and procedures which flow from these traditional constructs and assumptions. This sophistication in dealing with superficial matters creates the illusion that many of the discussions on standardized testing have considered basic issues. Therefore, a serious situation is presented for those who demand accountability from test designers who deal with implicit constructs and assumptions.

Among the significant, seldom debated, generaly unsupported, and widely held guild assumptions are the following. In one form or another, standardized test makers assume a degree of universalism in human behavior and experience among the world's people. Whether speaking of standardized tests of "intelligence," "personality," "achievement," "aptitude," or other areas, virtually all standardized test advocates assume a common starting point, a common rate of progress and common outcomes for efforts among the widely diverse cultural groups within America and even in the world as they approach problem solving situations. Because of this assumption, the idea of a "norm" has actually become an ideology among professionals in psychological assessment. In spite of all the criticisms of *standardization* which began almost as soon as the standdardized testing movement began (Lippman, 1976), there is as yet no serious attention in the literature of psychological assessment to demonstrate that specific superficial human behaviors can be demonstrated to be "normal" among widely diverse groups in America and even the world. Certainly the "norm" can be shown to exist *statistically* as an artifact of the measurement process. However, it is a continuing error in assessment that we confuse the *statistical* production of a *statistical* norm with the existence of a real *behavioral* norm. If these two assumptions are held of course, then the mass production of assessment instrumentation is possible, since individual questions for the interrogation of subjects can be standardized. It should be easy to see that mass production is more likely to lead to large scale profits than any likely alternative. In saying this, it must also be added that *appropriate* and valid standardized interrogation, if possible, would make better science than inappropriate and invalid standardized interrogation. The issue here is not whether standardized interrogation is desirable, but simply that if it is to be required as a basic element in the development of tests, then special consideration must be taken to insure the appropriateness of the standardization.

The basic problems with standardized testing can never be resolved until such time as most of the important constructs and assumptions are raised to a high level of professional consciousness and are examined freshly. It is at this point that we find great inertia in the field.

History justifies the critical concern that traditionally oppressed and excluded groups have had with standardized testing. Not only have these groups been the victims of incompetent science, there have also been times when racism in scholarship has been but thinly veiled (Carlson and Colburn, 1972; Chase, 1977; Friedman, 1970; Kamin, 1974; Lacy, 1973; Pearce, 1965; Stanton, 1969; Thomas and Sillen, 1972; Weinreich, 1946). African-Americans and other oppressed and excluded groups have been forced to sit mute and defenseless, and to observe the spectacle of strangers to their communities using presumably objective scientific devices for testing where the deck is stacked against African-American communities. This stacked deck has always been achieved in exactly the same way. It has been achieved by allowing interrogation of African-Americans to be based upon the cultural experiences of a small subset of White America. In general, leaders in the field of standardized testing have failed to deal with the substantive issue of test bias. In fact, some have even denied the existence of the issue of cultural bias, an African-American perspective on standardized testing can be expected to reflect multiple major concerns. In some ways, these concerns are the same as those which one would expect from any traditionally oppressed or excluded group. In still other ways, an Afro-American perspective on standardized assessment will reveal the need to deal in a systematic and scientific way with the unique cultural milieu within which a subject lives and from which all responses to interrogation must be fashioned. Assessors who fail here are simply ignorant of their own culture as well. These assessors commit the egocentric and immature fallacy of confusing their own culture with nature! (Hall, 1977)

The Existence of African and African-American Culture

Most standardized tests have been based upon the implicit assumption that the superficial world view, behaviors, and styles of a particular subset of white Americans would become the universal standard against which all other groups would be compared. We may think of "nature" and "culture" as being different in a very important way. Nature consists of those things which exist and which are not the deliberate creation of men and women. Culture, on the other hand, represents those things which are deliberately created by men and women out of the things which are available to them (Freire, 1972). The mental *functions* which allow these creativities to occur may be thought of as the "deep structures" of human mental behavior, such as the universal human ability to develop *a* language. However, the *particular* language which is developed is unique to a particular environment and may be thought of as the "surface structure." The overwhelming predisposition of European-American behavioral science has been to confuse these two levels of data. Therefore, for example, the only way that many Europeans or European-Americans have been able to observe Africans or African-Americans has been as a perceived deviation from a European or European-American "norm." I have referred to this elsewhere as a Type I question (Hilliard, 1976). It takes the form, "Do you know what I know?" Type I questions are the types of questions which characterize almost all standardized testing. A Type II question begins from an entirely different point. It takes the form, "What is it that you know?" Giants such as Jean Piaget are intellectually free enough to ask these. This forces any assessor, at least initially, to an analysis of human behavior that begins with the manifest culture of a particular people who may be involved. Following that, the probability for the discovery of a deep structure across human cultures is greatly enhanced, as Piaget's work has shown. For these reasons then, it is important that brief attention be paid to what should be evident to a sensitive observer, and that is the existence of African or African-American culture.

During Europe's colonial period, Africans came to be seen by Europeans as "Pagans," "uncivilized," "savages," "heathens," and so forth. In virtually all cases where these terms were used, the reasons for making such a determination were to highlight real or imagined differences between Europeans and Africans on European or European-American cultural behaviors. These differences could and did become the markers to target particular groups for oppression. It can be seen that by using such unscientific terms as those mentioned above, there is no likelihood, maybe even no possibility that real African or African-American culture can be observed, described, and seen as having an integrity of its own, different from, but equivalent to other cultural forms. Further, many European observers are destined to remain ignorant of their own culture as a culture. The gross mistakes by European observers of African culture which are due to this conceptual incarceration and cultural chauvinism are too numerous to recount here (Hilliard, 1976; Hall, 1977; Turner, 1949). However, we may cite a few examples. Emerging data from the systematic study of past and present Africa reveal, for example, that contrary to the published beliefs of most European scholars during the colonial period, African people were intensely religious. However, the religion which was lived fully by Africans was conceived and practiced in a form that European observers were unprepared to perceive, since their own beliefs (frequently irreligious) were used as a standard for comparison. The early studies by European linguists of African language, and especially by European-American linguists of African languages reveal that what has developed over the past few years has been a keener perception of African language by European and European-American linguists, and not a development of African language itself. In other words, African languages were, from the beginning, as sophisticated and as complex as European or any other languages anywhere in the world (Diop, 1978; Turner, 1949; Whorf, 1956). It was an unscientific and egocentric error to regard some languages as "primitive" and others as "developed." The immaturity, therefore, was not in the African language but in the European investigators' perceptions. The exact thing may be said of religion (Mbiti). Any student of African culture is continually amazed at the gross examples of the misassessment of African people on almost every count. While it may have served colonial European's interests to describe such things as African art as "primitive," a more sophisticated study of African art revealed this to have been another one of the typical gross errors in assessment.

Over and over again we find scholars, since the time of Europe's colonial period in Africa, revising upward their opinions on African culture. In recent years we have had outstanding descriptions of African culture as an integrated whole (Anyanwu, 1976; Diop, 1978; Jahn, 1961). These studies have been done by sensitive non-racist European and African scientists. The importance of this maturity and lack of racist bias in scholarship is that it allows a more scientific approach to an understanding of African people. Clearly, any scientific interrogation of African people is fruitless and senseless except when approached through African culture. In examining the emerging scholarship on African people, one becomes aware immediately of a general phenomenon among many European and European-American oriented scholars. That phenomenon is the resistance to any consideration of cultural data. Typical among the examples which might be cited is that of the outstanding scholar, Cheik A. Diop, who, as a part of his doctoral thesis which centered on his own culture, wrote what would later become regarded as a pivotal piece of scholarship on Africa. Diop had an uphill struggle to gain acceptance of his thesis of "cultural unity" within the continent of Africa in an era when European science seemed dedicated to proving the existence of wide diversity among Africans. Diop's struggle was not confined to his graduate committee in France, but extended also to scholars worldwide in his field who are themselves "experts" on Africa. Further, the fate of his book should be of interest. Here is a giant of a man who had a vital message, something which was eminently germane to issues which were being discussed in Europe and in Africa, and yet his book found its way only into limited circulation in English and, therefore, in America. Yet, every day the evidence is being amassed which supports that early thesis of Dr. Diop. The reception of this great scholar's work is like the reception of truth about African-Americans by many assessors who are alien to African-American communities.

The important thing to realize here is that the African culture which can be described today in all its richness was overlooked by some of the best scholars in the past who seemed oblivious to what was right before their eyes. This was in spite of the fact that prior to Europe's colonial period scholars had access to much data which could not conceivably have supported later self serving study of Africa.

It can be demonstrated that African-Americans, in general, participate in a cultural stream which is a resultant of both African and European-American experiences. A growing number of African, African-American, European and European-American scholars have been able to demonstrate conclusively the existence of significant African retention of culture within the general African-American population. This is of vital importance to the issue of standardized test development. For example, as Dr. Ernie Smith (1977) has shown, many Hamitic-Bantu or Niger-Congo languages morphologically do not utilize a copula to express "being". These languages utilize other rules for expressing the concept of being. Similarly, in these same languages, final consonant clusters are not used in the construction of words. Consequently, assessments of African-American language (Ebonics) which fail to recognize morphological, phonological, as well as syntactical, lexical, proxemics, semantics, stressing, and rhythm retentions from an African base are incompetent at their very foundation. Consequently, it would be pure gross error to say that an Ebonics speaker "deleted" a copula or "dropped" final consonant clusters, when these linguistic rules were never a part of the speaker's linguistic system in the first place. Naturally, this analysis says nothing about the practicality and perhaps the necessity for using both copula and final consonant clusters in America. This example only illustrates the fallacy of using such phenomena as a method for assessing or testing either the intelligence or the linguistic capabilities of Ebonics users among African-American children by using common English. Thousands of African-American children are referred for speech therapy in public schools simply because of their having learned a culturally appropriate way (Ebonics) of pronouncing words. Still others are assumed to be mentally deficient simply because the standardized interrogation cannot allow for cultural variation in the materials for competency demonstration. There is little evidence that standardized test makers and users are prepared to understand this.

Slavery and segregation (de facto and formal) in America are responsible in large part for the existence of African cultural retentions within the culture of the African-American. The isolation of African-Americans cause these retentions to be greater than otherwise would have been the case. Some of these retentions are so pervasive and so subtle that they are not recognized as retentions even by African-Americans themselves. It is only when sensitive students of African-American culture apply their skills to these areas that the truth can be known.

If African culture exists and if African-American culture exists, then the traditional implicit assumption of universalism and the traditional ideological assumption of "THE NORM" are impossible to accept on any rational basis at the surface structure level. The manifestations of culture in any group are reflected in its language, values, experiences (general information), beliefs, and general world-view. Among other things, each of these will modify the meaning of a client's response to interrogation. With diversity in the meaning of responses comes an immediate and fatal threat to the additive nature of those responses and, therefore, to the possibility for standardized processing (counting or computing). It is less important here to articulate further the descriptions of either African or African-American culture than it is to illustrate by the material above that the meaning of this material is ignored by those who construct standardized tests. It is not a matter of whether well trained scholars agree or disagree regarding the existence of an African-American culture. The problem is that most test makers are not sufficiently knowledgeable about the issue even to enter into the conversation. This is simply poor science. A much more detailed discussion of the African-American culture and assessment process is offered in my earlier work on I.Q. testing (Hilliard, 1976).

The Culture of the African and the European Projection

All assessment of human behavior is an interactive process. That is to say, some part of the outcome of any standardized test score or other form of assessment is due in part to the way a given subject responds. But the outcome or score is also due in part to the meaning that test construction people and interpreters give to the material and to other influences that they have on the outcome of an assessment process. Fortunately, in recent years there has been increasing systematic attention to the behavior of those who are engaged in the process of assessing people who are different from themselves, particularly assessors who enjoy a privileged position as a member of a privileged culture within a particular nation (Fanon, 1967; Hall, 1977). For example, many authors have done excellent studies and analyses of how the perceptions of dominated populations by members of a privileged class frequently represent little more than a projection onto a given population of the fantasies from those who do the studying (Curtin, 1973; Fanon, 1967; Jones, 1971; Pearce, 1965; Toll, 1974).

For example, we may look at the writings of European colonial historians about "savage" people, European-American critics of "low brow" music, European-American studies of "deprived families" and see unmistakable evidence that the European and European-American evaluators of Africans and African-Americans have most often told us more about themselves than about the people whom they purport to study. The contamination of constructs and assumptions by the projection onto traditionally oppressed and excluded groups of the visions of alien and privileged evaluators is clear. Once again, we have a case where the phenomenon (of projection) has been studied, but the outcomes of such a study have failed to penetrate the conservative perceptions of traditional scholars who develop standardized tests. These are not theoretical matters. The history of cross-cultural assessment will show that alien assessors frequently represent a major contaminating influence in most studies of human behavior causing the perpetuating of gross error. The typical response of standardized test makers to this problem is to ignore it out of existence.

Racism and Scholarship

A brief word must be written here regarding the unknown or unacknowledged role which some scholarship has played in the development of and perpetuation of racism. This racism in scholarship extends not only to the development of theories and misinformation, in general, but to the area of standardized test development, in particular. Without belaboring the point, it is sufficient to refer the reader to the excellent treatments of this subject by several authors whose documentation is thorough (Chase, 1977; Kamin, 1974; Stanton, 1969; Thomas and Sillen, 1972; Weinreich, 1946).The importance of this consideration is that it affects discussions about the necessity to guarantee adequate representation from all groups in the design and implementation of standardized tests. It will be to the everlasting discredit of American scholarship in testing, that even during the worst periods of segregation and racism in America, no systematic account of the possible impact of racism on the design and implementation of assessment appears. This represents a systematic bias of the highest order. Either we must conclude that racism was unimportant or irrelevant to the cross group assessment and, therefore, was unstudied, or we must conclude that it was vital and relevant but that our best minds either could not or would not tackle the problem. Since the problem is still with us, how can we we explain the continuing reticence of most test makers to tackle the issue?

Cultural Bias in Action: A Timely Example

There is a different set of concerns on the issue of cultural bias for every type of test. Therefore, any meaningful discussion of how cultural bias applies to standardized testing must take the types of tests or assessment issues one at a time. If this is not done, neither pro nor con arguments can have any rigor whatsoever. For example, standardized I.Q. tests are criticized because they use White common English for interrogation of many African-Americans who speak a form of Ebonics primarily. This means that the Ebonics speaker, unlike the White common English speaker, is really taking two tests instead of one! One is the test of whatever the problem or content test measures. The second test is a test of a knowledge of White common English. Consequently, the African-American Ebonics speaker is marked down on mental functioning for speaking a different language. When this matter is raised, a frequent response from defenders of the standardized test of I.Q goes something like this. "Well, Black people must learn to live and function in this society regardless of what their basic language is." Naturally, few sensible people could argue with this last conclusion for obviously practical reasons. The problem is that the conclusion is simply not a response to the criticism which was raised. This response is to an entirely different question. The absence of competence in White common English does not indicate the presence of mental impairment. Certainly it is useful and practical to know how to use White common English in America. But this is really a matter of teaching and learning and not a matter of medical therapy. Perhaps the ability to stay on the subject being raised should be offered as a new indicator of intelligence.

Examples such as the above happen all the time. African-Americans who must suffer from the arbitrariness and midstream shifting of focus of standardized test makers and users come to expect the common world of scientific absurdity. African-Americans have come to expect the "rubber-ruler" in "standardized" testing. When the standard question is actually measuring entirely different things in two different groups, a distribution of such scores represents the height of folly and ignorance masquerading as science.

Let's move closer to a classical case of absurdity in action, the National Teachers Examination. This standardized test already affects the lives of hundreds of thousands of African-American children and tens of thousands of African-American teachers as well as other teacher and student populations. Decisions about hiring, promotion, pay, and perhaps even the assignment of teachers are made on the basis of this test. It bears closer scrutiny because its use has already resulted in the denial of work to many African-American teachers and has legitimized the "competence" of many teachers who have not been able to produce educational gains for African-American children.

The Educational Testing Service reports that twenty-one states now require some part or all of the NTE according to ETS Developments, 1978 (Number 2). Because of low scores on this examination, the number of African-American teachers in South Carolina and in California have been drastically reduced. At the same time, there is no evidence that the school performance of African-American children in these two states has improved as a consequence of their being taught by teachers who score high on the NTE. Yet, the federal district court in South Carolina, and the Supreme Court of the United States, ruled that the NTE could continue to be used, since South Carolina did not "intend" to discriminate in its use of the NTE, and "since we find that the NTE creates classification only on permissable bases (presence or absence of knowledge or skills and ability in applying knowledge), and that they are not used pursuant to any inten to discriminate, their use in making certification decisions by the State is proper and legal" (ETS Developments, 1978). I was not aware that the validity of a test could be established scientifically by a legal decision regarding the intention of users. The citation of this legal decision tends to imply a kind of validity for the NTE. Following this decision, an African-American employee of ETS, William Harris, whose picture is shown in the article is quoted as saying:

> Test questions are carefuly reviewed to eliminate possible negative bias against minorities...the fact that some minority candiates score substantially lower is a reflection, not of test bias, but of educational difficulties that minority students have experienced.

It is not clear from the article just what magic wand Mr. Harris or other ETS personnel were able to wave in order to "eliminate the negative bias against minorities".

It gets worse. ETS reports that a "content validity" study was conducted on the NTE. This study is similar to one presently being conducted in California. The "content validity" of the test is determined by surveying 456 South Carolina college and university faculty, who in this opinion poll gave testimonials which "found the NTE to be a fair measure of the knowledge imparted by the South Carolina teacher training programs." What kind of validity check is this? When did face validity, even by faculty, become so scientifically acceptable that it could be reported with a straight face? Then, in an incredible statement in ETS Developments, ETS declares that the NTE, "are also not intended to be a measure of classroom teaching performance, but rather of the academic knowledge basic to successful teaching" (emphasis mine). How does ETS determine "the academic knowledge basic to successful teaching?" How is successful teaching considered not to be a matter of performance?

It's hard to believe that the professional careers of so many teachers could be tied to such hocus pocus. From an African-American perspective, at least, the purpose of a teaching certificate is to certify a quality performer. There is no earthly reason for certification if it makes no difference for the client by insuring that the client is protected from professionals who cannot perform. It seems to me that the gross problems with the NTE and its use include the following:

1. There is an implicit assumption that there is a standard curriculum for teacher preparation nationwide. No such curriculum exists, nor is there any serious attempt to demonstrate that it does. It is ludicrous to assume that 456 or any other number of South Carolina college and university faculty have done the systematic homework to know if the test content matched South Carolina teacher education content. Who has the data to answer that question for the whole nation? That could be determined only by a systematic examination of course syllabi, tests, reference lists, assignments, and observation of instruction. This has not been done. Further, there is reason to expect wide diversity in curricula offerings. I have spent most of my professional life in teacher education and am totally surprised that ETS seems to have discovered a universal teacher education curriculum, either academic or professional. In fact, the whole

1978 Hunt Lecture at the American Association of Colleges for Teacher Education by Dr. Lawrence Cremin of Columbia University documented almost the exact opoposite! The professional preparation of teacher educators was shown to be almost totally ad hoc. Therefore, the instruction which they offer is probably ad hoc as well. At the very least, we can say that there is no universal content. Opinion polls and testimonials may be good advertising but certainly cannot be considered good science.

2. There is an implicit assumption that the knowledge of what the NTE measure is related to the curricula of primary and secondary schools and to successful teaching practice. We have virtually no data on this.

3. There may also be an implicit assumption that there ought to be a common content for the teaching bachelor's degree. This is wild in view of the fact that there are no attempts anywhere in academe, to my knowledge, to make a universal curriculum for higher education. The academy, in the name of academic freedom, has demanded almost the opposite.

4. There seems also to be an implicit assumption that if there were a common or universal content that there would be no "bias" against African-Americans. The whole challenge of the Black Studies movement and other ethnic studies movements calls that very matter into question. Clearly, there has been a permeating European-American ethnocentrism in the curricula of American education. A review of the NTE reveals that the same bias is still present in the examination itself, the ETS "careful review to eliminate possible negative bias against minorities" notwithstanding. The higher education curriculum has yet to be developed which rectifies over four hundred years of academic bias in scholarship. Therefore, the ETS test is unlikely to have achieved that freedom from bias, especially since it is measuring its version of universal curriculum.

Where Do We Go from Here?

If serious attention is to be devoted to the elimination of bias in standardized testing, and if scientifically sound and culturally appropriate testing is to be developed, an effort of grand proportions will be required. To be fair, accurate, and competent in the assessment of African-American populations the following things will be required.

1. If mental functions are to be assessed, the functions to be assessed must first be identified with precision. Following that, appropriate culturally specific materials will be used for interrogation and the results will be interpreted by culturally sensitive assessors. For example, if the mental function of "conservation" is to be assessed, it can be seen that a variety of vocabulary and interrogation styles can be used to discover its use by a client. In other words, the assessors must be sufficiently sophisticated to separate the function (deep structure) to be assessed from the possible cultural manifestation of that function (surface structure). This is analogous to Noam Chomsky's formulation in linguistics.

2. If mental functions are to be assessed in African-American populations, it should happen only after extensive multidisciplinary review of the assessment process by culturally sensitive professionals. For example, since language is a primary tool for interrogation, and since language is asserted as a reflection of mental functions, then standardized assessment of African-American populations should never be done unless the procedures can pass the evaluative scrutiny of experts on African-American language. Similarly, since there is a growing literature on culture and perception, and culture and world view (Cole, 1974; Levi-Straus, 1962; Nobles, 1976; King, 1975), no standardized testing of African-American mental functions should be attempted unless the procedures can pass the evaluative scrutiny of culturally sensitive experts on African-American culture. Clearly, expertise in the mechanics of test construction in no way qualifies such experts to interpret African-American responses to interrogation.

3. If the "achievement" of African-Americans is to be assessed, one of two conditions must obtain. The criterion of achievement must be demonstrated to be other than an *arbitrary* standard which bears no valid connection to screening or status evaluation goals. For example, most of the NTE content is *arbitrarily* determined. Further, African-Americans have the right as citizens and the human right not to be required to become "expert achievers" in test content which may well be biased against African-Americans. For example, a test of general information and history which failed to treat the activities of Toussaint, L'Ouveture, Marcus Garvey, Edward Wilmont Blyden, Carter G. Woodson, Paul Robeson, Ira Aldredge, and others who have been heroic in the struggle for the liberation of African-Americans, would represent gross misinformation for African-American people. In reading the questions on the *NTE*, the *College Boards*, the *Iowa Tests of Educational Developments*, and other standardized tests of achievement, one would get the impression that African-Americans were non-existent and that the condition of slavery and other forms of oppression was unimportant for an educated person. Indeed, to be judged as a competent teacher by the *NTE*, in South Carolina no less, or elsewhere, some African-Americans would have to commit cultural suicide or at least to regard long and intensive study to obtain a knowledge of self and group as a total waste of time. An African-American could be highly knowledgeable about the history of Africa, African-Americans in America, and hold a highly sophisticated view of world and American history, but would receive little or no credit for that knowledge simply because that knowledge was either unknown to, or unused by the test developers. Therefore, if achievement is to be measured by standardized tests, then African-Americans have every right to have those tests evaluated by interdisciplinary experts who are knowledgeable about essential African-American oriented content in each academic area. For example, it is of great interest to much maligned African-Americans to know that most of the Greek scholars who are regarded as the founders of "Western Civilization" received their higher education in great universities in Africa and were taught by Africans (James, 1976; Fraser, 1972). No achievement tests, to my knowledge, deal with such matters at all. In fact, one wonders if the "achievements" of the item writers is sufficient for them to deal with such material.

4. If the errors of standardized testing of African-American populations are to be rectified, the band-aid approach to the solution of the problem will have to be abandoned. For example, the cultural bias of a test cannot be corrected simply by changing the number of African-Americans in the standardization sample, as was done on the revised Wechsler Intelligence Scale for Children. The problem of cultural bias is neither in the counting nor in the sample but in the items. The items must be culturally appropriate. It's time to be serious.

5. In the mad rush to respond to charges of cultural bias, there has been a rash of "alternative" tests developed. For example, Dr. Jane Mercer has done an outstanding job of illuminating the weaknesses of standardized I.Q. tests. She was most helpful in demonstrating the culture-bound nature of assessment processes. However, Dr. Mercer has now succumbed to the temptation to put forth her own standardized assessment system (System of Pluralistic Multicultural Assessment, SOMPA). Incredibly, it contains some of the same tests which were under attack. Further, no respect at all was shown for African-American or other cultural data in the SOMPA system. It could not, since SOMPA is all things to all people. It appears to have all the weaknesses of the old tests, plus a whole host of new weaknesses all its own, not the least of which is the absence of any construct validity. In fact, there does not even appear to be a construct to validate. "Adaptive behavior" is so poorly defined that it cannot satisfy the demands for assessment rigor. The alternative to a test which did not measure what was intended may well be nothing, if there was nothing there to be measured in the first place. I can see that SOMPA is a system of assessment. I do not know what in the world the results would be good for. The lure of standardization is just irresistible for many assessors.

6. If the errors of standardized testing with African-American populations are to be corrected, test makers and users must be able to make clear distinctions between research and evaluation. Most standardized tests should be regarded as experimental tools. Since there is no "pure food and drug administration" for standardized tests, the public is at the mercy of the judgment of the manufacturers who must be trusted to protect both their own self interest and that of the public as well. It will not work. A good salesperson can sell, and has sold, the most inadequate (but beautifully packaged) instruments to an unprotected public. To leap from experimentation to everyday application should require a powerful demonstration of validity and the absence of harmful side effects. Clearly, this has not been the case with standardized testing and African-American populations.

7. If standardized tests are to be improved, the makers and users of those tests must be taught the difference between "race" and culture. Otherwise, esoteric sounding academic discussions will be nothing more than nonsense. For example, if the Weschler Intelligence Scale for Children is to use a sample stratified by "race," how in the world are the subjects to be selected and assigned to categories? In other words, what are the scientific (genetic, physical, biological) criteria for identifying the subjects by race for the sample? A driver's license? A confession? The examiner's observation? Anthropologists gave up the concept of race long ago (Montagu, 1974). What is the meaning for psychologists?

Culture, on the other hand, does exist, but how is it to be defined and treated? We are just beginning and should shed all pomposity about cross-cultural assessment until we have something to offer.

Conclusion

Standardized testing has a bad heredity to overcome. It must overcome the ignorance of culture by its prime movers. It must overcome both overt and covert racism among many of its supporters. It must overcome the reluctance of its prime movers to deal adequately with the question of validity and utility of testing for improving the human condition.

The mathematical treatment of things that do not even exist and the scaling of scores on items where (in cross cultural settings) meaning does not "commute" is an exercise in self deception. The items in most standardized tests of "ability" do not meet scientific criteria for aggregation! The real issue before us is to decide if the repair of standardized testing is to be but an exercise in cosmetology or if it is to be a searching full scale troubleshooting job. We have the professional skills and tools to do the job that is necessary. The only reason that I am placed in the position of writing a position paper on African-Americans and standardized testing is that the things that we know to do about testing have not been done. Some of the things that we must do to remedy the situation will create more problems in terms of politics than professionalism. For example, when we stop labeling African-American boys as mentally retarded when they are nowhere near that, then they will be returned to regular school environments where many are unwelcome because of things that have nothing to do with their mentality. Similarly, when achievement test content is changed to reflect critical content for all Americans, someone must be prepared to accept the flack from those who enjoy building their egos at the expense of others. However, in the end, honest, humane, and professional psychologists and educators will be able to face the consequences of their work and still sleep like a log every night.

References

Anyanwu, C.K. The Nature of Black Cultural Reality. Washington, D.C.: University Press of America, 1976.

Carlson, L.H. and G.A. Colburn. In their place: white America defines her minorities. New York: John Wiley and Sons, 1972.

Chase, Allan. The Legacy of Malthus: The Social Costs of New Scientific Racism. New York: A. Knoph, 1977.

Cole, M. and S. Scribner. Culture and Thought. New York: John Wiley, 1974.

Curtin, P.D. The Image of Africa: British Ideas and Action, 1780-1850, Madison: University of Wisconsin Press, 1973.

Diop, C.A. The Cultural Unity of Black Africa. Chicago: Third World Press, 1978.

Fanon, F. A Dying Colonialism. New York: Grove Press, 1967.

Fraser, P.M. Ptolemaic Alexandria. Oxford: Clarendon, 1972.

Freire, Paulo. Pedagogy of the Oppressed. New York: Herder and Herder, 1972.

Friedman, L.J. The White Savage: Racial Fantasies in the Postbellum South. Englewood Cliffs, N.J.: Prentice Hall, 1970.

Hall, E.T. Beyond Culture. New York: Anchor, 1977.

Hammond, Dorothy and Alta Jablow. The Myth of Africa. New York: The Libary of Social Science, 1977.

Hilliard, Asa G. III. Alternatives to I.Q. Testing: An Approach to the Identification of Gifted "Minority" Children. ED 145-957, 1976.

Hilliard, Asa G. III. "Standardization and Cultural Bias as Impediments to the Scientific Study of Intelligence". Journal of Research and Development in Education, 12, 2, 1979, pp. 47-58.

Hilliard, Asa G. III. "The Predictive Validity of Norm Referenced Standardized Tests". Negro Educational Review, 28, 3 & 4, 1977, pp. 189-201.

Jahn, J. Muntu: The New African Culture. New York: Grove Press, 1961.

James, G.G.M. Stolen Legacy. San Francisco: Julian Richardson, 1976.

Jones, Eldred D. The Elizabethan Image of Africa. University Press of Virginia, 1971.

Kamin, Leon. The Science and Politics of I.Q. New York: John Wiley and Sons, 1974.

King, L. African Philosohy: Assumptions and Paradigms for Research on Black Persons. NIMH Grant ROLMH 255580-01, Los Angeles: Fanon Institute, 1975.

Lacy, Dan. The White Use of Blacks in America: 350 years of law and violence, attitudes and etiquette, politics and change. New York: McGraw Hill, 1973.

Levi-Strauss, Claude. The Savage Mind. Chicago: University of Chicago Press, 1962.

Lippman, Walter. "The Lippman-Terman Debate: The Mental Age of Americans". In N.J. Block and G. Dworkin (Eds.) The I.Q. Controversy. New York: Random House, 1976.

Montagu, A. Man's most dangerous Myth: The fallacy of Race. Oxford University Press, 1974.

Nobles, W. "Psychological Research on the Black Self Concept." J. Social Issues, 29, 1, 11-30, 1976.

Pearce, R.H. Savagism and Civilization. Baltimore: The John Hopkins Press, 1965.

Smith, Ernie. Keynote Address to Black Child Development Institute Annual Meeting, Houston, 1977. (Professor of Linguistics, California State University, Fullerton).

Smith, Ernie A. "The Retention of the Phonological, Phonemic and Morpho phonemic Features of Africa in Afro-American Ebonics". Seminar Series Paper, Department of Linguistics, February 1978. California State University, Fullerton.

Stanton, William. The Leopard's Spots: Scientific Attitudes Toward Race in America. 1815-19, Chicago: University of Chicago Press, 1969.

Thomas, A. and S. Sillen. Racism and Psychiatry. New York: Brunner/Mazel, 1972.

Turner, Lorenzo. Africanisms in the Gullah Dialect. Chicago: University of Chicago Press, 1949.

Weinreich, M. Hitler's Professors: The Part of Scholarship in Germany's Crimes Against the Jewish People. New York: YIVO, 1946.

Whorf, Benjamin. Language, Thought, and Reality. Cambridge: The MIT Press, 1956.

CRITICAL FACTORS IN THE TESTING OF HISPANIC AMERICANS: A REVIEW AND SOME SUGGESTIONS FOR THE FUTURE

Amado M. Padilla
Department of Psychology
University of California, Los Angeles

Abstract

The literature on testing of Hispanics is reviewed. The review reveals that a number of factors have been identified that influence test performance of Hispanics. The major factors are related to: language, acculturation, and socioeconomic status. A correction method for use with Hispanics which takes into consideration these factors was suggested over 40 years ago and has recently been implemented. It is argued that any testing program must be cognizant of these factors. Recommendations are that test makers and consumers should: (a) employ culturally sensitive testing procedures, (b) design and construct methods of assessment that are criterion referenced, and (c) select models of culture-fair assessment that are based on utility theoretic approaches.

Testing has become a major industry in the United Stated (Kohn, 1975). There is probably no person today who has not been affected in some way by standardized tests. Houts (1975) has estimated, for example, that in the course of a child's schooling, from kindergarten through high school, he or she will take between six to twelve full batteries of achievement tests. In addition, the child may be tested on a number of other standardized instruments ranging from reading readiness to IQ tests. Still later individuals are required to take standardized tests to gain admission to a university, graduate or professional school, or some other vocational program. A large number of other aptitude tests may also be required along the line for career selection and advancement. Our society can truly be called the Age of Testing.

Testing, as an industry, has been criticized repeatedly over the last two decades (e.g., Hoffmann, 1964; Houts, 1977). In spite of the criticisms levied against testing, few critics have seriously campaigned for the total elimination of standardized tests. The anti-testing advocates have usually argued that in their current form tests possess little predictive validity and have suggested ways in which tests could be improved and abuses due to testing eliminated. Among the arguments heard against current standardized tests is that they are biased against certain racial/ethnic groups. Garcia (1977) has gone so far as to suggest that there is so much quackery involved in tests because of the way in which they were constructed that they should not be taken seriously by anyone, least of all when they are used with ethnic group members who differ culturally and linguistically from the majority group on which the tests were standardized.

In this article, we will examine testing from the perspective of the Hispanic American. Our intent will be to review the literatue on testing. Secondly, on the basis of the review, we hope to offer a series of recommendations for improved use of tests among Hispanics. We do not pretend that these recommendations will be novel or startling. Many have been in the literature approximately 45 years but bear repeating. The important consideration is that for as long as many recommendations have been in the literature failure to take these suggestions seriously by tests makers and by the consumers of tests has resulted in a situation wherein the use of tests with Hispanic Americans must be called into question.

Before proceeding to a review of literature, it is important to note that the term Hispanic is used here in its generic sense to refer to people of Spanish origin who reside in the United States and the Commonwealth of Puerto Rico. Specifically, we are referring to people of Mexican, Cuban, Puerto Rican, and other Central or South American ancestry. Population estimates for this group range from the 11 million reported in the 1970 census to more recent estimates of up to 20 million. The figures do not include the 3.2 million inhabitants of the Commonwealth of Puerto Rico. Accordingly, by Hispanic we are referring to between 7 and 12 percent of the United States' general population. As a final note here, it is important to state that in the review that follows we have maintained the term used by the author(s) to describe their subject population (e.g., Mexican, Puerto Rican, etc.). The reader will thus observe that most of the literature is confined to the largest Hispanic subgroup, that is, Mexican Americans, and aside from recent studies with Puerto Rican subjects, studies using other Hispanic subgroups are notably absent. Therefore, even though the recommendations that follow the review are based on research with at best two Hispanic subgroups, the generality of the suggestions across all Hispanic subgroups is supportable.

Finally, the review of literature that follows is almost entirely devoted to articles on intelligence testing. This is not deliberate, but is an accurate reflection of the literature. Until recently, few studies have been conducted using tests other than standardized IQ measures.

The Early Period of Testing: 1923 to 1940

Garth (1923) presents a study where among other hypotheses, he proposes to test "...the principle that mixtures of different lines brings about differences in intelligence" (p. 388). To evaluate this assertion, he administered the National Intelligence Test to five "blood groups" representing varying degrees of white and Indian blood. One of the groups tested "represented the typical 'Mexican' whose ancestry is largely of Spanish blood and of various ... nomadic tribes of Mexico" (p. 389). Although Garth's data do not warrant extended citation, a few quotations are included to identify racist attitudes which characterized the efforts of some workers during that period. Garth states that because "social status and education have not been controlled, we may not positively state that these data indicate racial differences in intelligence" (p. 401). Nevertheless, the tenor of his position is revealed by a number of unqualified statements on intelligance, such as,

...mixed breeds excel the pure breeds in intelligence scores

> ...If these groups may be taken as representative of their racial stocks, the results indicate differences between their racial stocks in intelligence as here measured ... one is inclined to believe that differences in ... mental attitude toward the white man's way of thinking and living are here made apparent. (p. 401)

In a second article on the intelligence of "Mexican" children, Sheldon (1924) recognized the problem of testing bilingual children:

> ...one school in this city was found to be devoted entirely to Mexicans, and here the children understood Engligh very imperfectly. To overcome the difficulty ... it was necessary that the children ... be given the group tests by their respective teachers. These teachers were able to make themselves understood by the use of a sort of Spanish-English dialect colloquially called "spic", or a mongrel Spanish. (p. 140)

Despite the statement above, it is clear that Sheldon fails to recognize that this type of language difficulty in and of itself, will depress IQ scores. Thus, Sheldon concluded that the average Mexican child in the first grade was 14 months below the normal mental development for white children of the same school grade.

In a lengthy monograph on the racial differences in the mental and physical development of Mexican children, Paschal and Sullivan (1925) noted that a simple translation of the Binet test into Spanish was not adequate "owing to the different connotations of supposedly equivalent words" (p.6). Accordingly, they tested all of the Mexican 9 and 12-year-old children in the Tucson, Arizona, schools with six different performance scales. Upon comparing the mental test scores of Mexican children with the American norms, considerable overlap was found between the two groups for the top third of the distribution. But from the median downward the American children showed an increasing superiority. Perhaps more interesting was the fact that the difference between the two groups was greater at 9 years than 12. Paschal and Sullivan speculate that the greater difference between American children and the Mexican children at 9 years of age may be indicative of the relatively greater part played by the school in the mental development of the latter group.

Garth (1928) tested 1,004 Mexican children and compared their test scores with normative data from the National Intelligence Test. Lower mean scores support normative data from the National Intelligence Test. Lower mean scores support earlier findings of Sheldon. But in this article Garth does not acknowledge that the scores may be deflated because of language difficulties.

Two studies appeared in the same year using a draw-a-man test. In the first, Manuel and Hughes (1932) found that at each age level between 7 and 10 years, Mexican children have mental age scores approximately 1 year lower than expected. In the second investigation, Davenport (1932) compared younger and older siblings of both Mexican and non-Mexican families. Mexicans scored lower than the non-Mexicans, but interestingly the older Mexican subgroup obtained higher IQ scores than their younger siblings.

These results were interpreted by Davenport to mean that the longer school attendance of the older Mexican children had significantly increased their level of intellectual functioning. Accordingly, Davenport cautioned:

> These findings, taken as a whole, point strongly to the conclusion that prognosis on the basis of the Goodenough IQ at school entrance is likely to underestimate the future abilities of the Mexican child. (p. 306)

In a subsequent report, Garth and Johnson (1934) argue that "Mexican children in the United States" are more like the American white (i.e., based on normative data) on the Otis Classification Test at the early ages, but less like them as they grow older in intelligence and achievement" (p. 228).

In a fourth study (Garth, Elson, and Morton, 1936), it is reported that Mexican Americans are consistently inferior to white norms in verbal intelligence and achievement scores. They describe "rather startling" data which indicate Mexican Americans do as well as American whites on a non-verbal measure, the Pintner Non-Language Intelligence Test. They understate the obvious conclusion that "Such results suggest that the Mexican child in the U.S. is handicapped when he is tested with a verbal intelligence test" (p.55).

Garretson's 1928 study illustrates how bias may lead to misinterpretation of data. He cites intelligence test data from children who are residents of a small Arizona town in which the major industry is copper smelting. He reports that the heads of Mexican households earn approximately $125 - $150 per month and that 25 percent are high school graduates. By comparison, "no more than 80 percent" of the heads of Mexican households are employed and their monthly income is estimated as $110 per month. He also observes that it is "unusual" for a Mexican child to enter school with any knowledge of English. Despite these overwhelming socioeconomic and linguistic differences, he interprets lower group mean intelligence test scores among this group as confirming some type of deficiency in genetic endowment; for example, "the native capacity, if we assume that intelligence quotients are indicative of native capacity, of the Mexican pupil is less than that of the American child" (p. 38).

In a series of 45-year-old papers which retain their relevance even today, Sanchez (1932a and b; and 1934a and b) rebuts the interpretations of genetic deficiency with a combination of experimental research, demographic data, and theoretical argument. In the two early papers he identifies five factors which could underlie lower mean IQ scores among Mexican children and evaluates each; "(a) hereditary limitations, (b) inferior home environment, (c) language handicaps, (d) unsuitability of tests, and (e) lack of parallelism of conditions under which tests are given" (1932a, p. 223). A term used by Sanchez ("(b) inferior home environment") must be defined to minimize possible misunderstanding of what he meant. Both 1932 papers indicate clearly that this term refers only to socio-economic deprivation and educational limitations (see Sanchez 1932a, p. 224 and 1932b, p. 551). Sanchez rejects the principle of racial inferiority because of the absence of any corroborative evidence. Furthermore, Sanchez

(1934) demonstrates significant increases in mean scores derived from tests of intelligence and reading ability when alternate forms of these tests are readministered on four occasions over a period of approximately 16 months. Although the absence of a control group weakens his position, he concludes that the first test score elicited from these children represents an underestimate and that factors (b) through (e) comprise the most reasonable explanation for this inaccuracy.

In his two 1934 papers, Sanchez makes several interrelated points. Even after 1 year of school, the average Mexican child is deficient in English vocabulary. Sanchez identifies some of these words and lists the frequency with which they appear on the 1929 Stanford Revision of the Binet-Simon tests. He also discusses how homonyms and word-usage differences in English can be confusing to the Spanish-speaking child. For instance, Sanchez (1934a) states:

> Such words as "like," "right," "kind," "get," and "look," used in the tests and commonly used with at least two different meanings, present serious difficulties to the child just acquiring a new vocabulary. Over and above this obstacle is the more nebulous, and consequently more trying, problem of the equivalence of concepts and of organized ideas ... "picture" connotes retrato (portrait) to the Spanish-speaking child and ... it is far fetched to call a drawing such as those used in the Binet tests a retrato. In a similar manner, "What is your name?" is expressed in Spanish by "What do you call yourself?" (Como te llamas?), while "How old are you?" is expressed "How many years do you have?" (Cuantos anos tienes?) etc. (p. 400-401)

He criticizes the use of Spanish translations of tests since these were being applied without independent validation. His most telling criticism concerns the "failure ... of the public school system" to provide "the Spanish speaking group ... with a comparable education" (1932b, p. 769). He presents data indicating that in New Mexico, during the school year 1932-1933, there were almost 25,000 Spanish speaking children enrolled in the first and second grades, but only 540 in the final year of high school. His point is that 2 percent of the Hispanic children got this far in school, compared to 14 percent of children from all other ethnic groups in the same State during the same period.

It is imperative that the reader grasp the full implications of the work of Sanchez. First of all, there is the man himself. He is Hispanic and familiar with all the nuances of his own culture -- language, family process, and socio-economic conditions. Second, his education provided him with skill in the generation and interpretation of social science data bearing on achievement and education. Thus, he is qualified to provide significant insight into research findings bearing upon Hispanics. Yet, although he identified problems and suggested resolutions more than 45 years ago, very little ameliorative work has been done. With the possible exception of the standardization of the WAIS in Puerto Rico (Green 1964 and 1969), little effort has been expended on standardizing a test of intelligence for Hispanics.

But Sanchez' conclusions are not the only ones which have gone unheeded. In two very similar articles Mitchell (1937) and Mahakian (1938) test bilingual children on two forms of the Otis Group Intelligence Scale. A counterbalanced design was used by both investigators such that half of the children were administered the Otis first in English and then in Spanish; the other half of the children had the language order reversed. Both investigators reported that the bilingual children did significantly better when tested in Spanish rather than English. Mitchell and Mahakian report mean differences between the English and Spanish forms of 13.22 and 7.6 points with a range of 44 and 33 points, respectively.

Both authors conclude that intelligence tests administered in English to Spanish-speaking children are not valid. Numerous recommendations were offered by Mitchell and Mahakian for the proper assessment of intelligence among these children. Mitchell, for example, suggested that a corrective factor be added to the intelligence test score of a Spanish language dominant child if tested in English to reflect more accurately his true level of intellectual functioning.

Of importance here is that some of the studies reviewed represent beginning attempts to identify with greater precision those environmental variables which depress scores on tests of intelligence. Even then, there was some limited recognition that IQ scores can be depressed by social, economic, and linguistic factors. Basically, much of the research published on testing with Hispanics from about 1940 on represents attempts to measure the relative influence of these and other variables upon test scores.

Testing from 1940 to 1960

It is difficult to discern patterns from the scattered literature published between 1940 and 1960. Most authors try to deal simultaneously with a large number of confounded environmental variables, there are few instances of programmatic research, and the literature has little heuristic value. Thus, it will be necessary to skip from article to article and from issue to issue.

Shotwell (1945) compared the performance of 80 "Mexican" and "American" mental defectives on verbal (Stanford-Binet) and nonverbal (Arthur Performance Scale) tests of intelligence. Both groups scored higher on the Arthur than on the Binet test, but the Mexican mean was 22 points higher. To explain the obtained difference in test scores, Shotwell entertained two hypotheses. First, the Mexican American may be unduly penalized when intelligence is measured by verbal tests alone. The alternative hypothesis is that:

> It may be that Mexicans of a moron or borderline grade of intelligance have a special aptness along manual lines that is not found in Americans who are similarly retarded mentally. (p. 448)

Shotwell concludes by suggesting that additional research is needed before deciding between the two hypotheses.

One of the earliest reported attempts to enhance reading readiness among Hispanic children of preschool age was published by Herr (1946). In this study a group of children who participated in a preschool program was compared with a control group of children who did not. Major results indicated that the mean IQ score (Pintner-Cunningham Primary Intelligence Test), of the children who participated increased almost 30 points whereas the mean IQ of the control subjects increased less than 10 points. These data, and much more than can be reported in a review of this type, indicate the "...pre-first-grade training is an important factor in success in learning to read among Spanish-American children" (p.102). In terms of recommending programs of social change, we will reintroduce this study with its conclusions and implications in a later part of this article.

Altus (1945) describes an Army program designed to teach reading and arithmetic skills to six groups of functional illiterates including five different ethnic minority groups and one group of whites. The lowest correlations between Wechsler-Bellevue subtest scores and performance in the programs are found among Americans of Mexican descent. Altus observed that some subjects with scores in the "feeble minded" range not only succeeded in the program, but also had work histories indicating a successful life adjustment. He concludes, therefore, that this particular test of intelligence has minimal validity of prediction of these kinds of behavior (school and work) among members of this group. Altus' conclusion that different ethnic groups perform differentially as a function of the test used is confirmed by additional work on individual and group tests of intelligence, as well as by a measure of mechanical ability (Altus 1948).

In yet another article, Altus (1949) reports that:

> ...of the personnel cards for all White trainees in the Center, it was found that exactly 48.1 percent of such trainees were either bilingual Americans of Mexican ancestry, speaking both Spanish and English; or non-English, Spanish-speaking trainees, mainly native to Mexico though a few came from the Central American countries. (p. 211)

In analysis of four performance subtests of the Army Wechsler Scale (now known as Wechsler-Bellevue, II), Altus' data indicated that the Mexican who could not speak English -- even if he had been born in the United States and had lived there all his life -- was on the average duller than the foreign-born, recently arrived Hispanics. Although Altus does not attempt to explain this rather interesting finding, he does speculate on the persistence of cultural and linguistic factors among American citizens of Mexican ancestry.

The sole attempt to establish norms based on a Hispanic sample was provided by Ammons and Aguero (1950). These investigators tested four males and four females from grades 1 to 10 and ages 7 to 16 on the Full Range Picture Vocabulary test (FRPV) and on the vocabulary subtest of the 1937 Stanford-Binet. Forms A and B of the FRPV correlated .86 with each other, and .85 and .82 with the Binet vocabulary test. Of significance here was the finding that as subjects grow older, the discrepancy between scores on verbal and nonverbal tests also increases.

Carlson and Henderson (1950) argue that most studies comparing intelligence among different ethnic groups ignore too many factors which influence IQ. They identify: (a) rural vs. urban environment, (b) socioeconomic level, (c) total cultural complex, (d) quantity and quality of formal education of parents and students, (e) diet, (f) examiner prejudice, (g) motivation, and (h) bilingualism. Although they cite Sanchez (1934a), they fail to acknowledge that he made many of these points 16 years before. After an unsuccessful attempt to control as many of these variables as possible, the authors compared IQ scores of Mexican and non-Mexican children who were 5 1/2 years and older. Unfortunately this study is inconclusive since, as Carlson and Henderson themselves acknowledge, they were unable to control for rural-urban background of parents, vocabulary level in English and Spanish, and a possible difference in motivation level between the two groups. Pointed criticisms of this study were made by Pasamanick (1951).

Cook and Arthur (1951) obtained mean IQ scores of 84 on the Binet and of 101 on the Arthur Point Scale of Performance from 94 migrant Mexican-American children in the area of St. Paul, Minnesota. Since this difference is highly significant, these authors argue that the Binet test is inappropriate and that low IQ scores do not, therefore, necessarily imply general mental retardation. The findings clearly point to the fact that there is a high degree of potential ability not tapped by the Binet test.

The earliest study using Puerto Rican subjects found a similar difference in IQ scores among this group of children when tested on the Pintner Verbal Test and the Pintner Non-Language Test (Darcy 1952). A mean difference of 8.28 points in favor of the nonlanguage test led Darcy to conclude that when testing bilingual children both verbal and nonlanguage scales must be used to obtain a valid estimate of the intelligence among the bilingual.

Although all the studies reviewed thus far have involved bilingual subjects, the first study in which degree of bilingualism was assessed was reported by Johnson (1953). In this investigation, the "Hoffman Bilingual Schedule" and a "Reaction-Time Technique" were used to measure bilingualism among 30 boys between the ages of 9 and 12 years. Intelligence was measured by the Goodenough Draw-a-Man Test and the Otis Test of Mental Ability. The findings indicated a negative relationship between performance on the Otis, a verbal test of intelligence, and degree of bilingualism; more bilingualism was associated with lower IQ scores on the Otis test. On the other hand, degree of bilingualism was positively correlated with scores on the Goodenough test. Moreover, the discrepancy between the Otis and the Goodenough IQ scores was least among subjects who evidence greater knowledge of English than of Spanish. Although earlier investigators had suggested that degree of bilingualism was correlated negatively with verbal IQ, this is the first study to actually document the relationship empirically.

The degree of bilingualism and its relationship to IQ scores is also discussed by Anastasi in two studies with children in New York City (Anastasi and Cordova 1953 and Anastasi and de Jesus 1953). Counterbalancing administration of the Cattell Culture Free Intelligence Test in English or Spanish, Anastasi and Cordova report that a marked improvement occurs from first to second testing, regardless of language. It was also found that girls performed better when the testing order was Spanish-English, the boys when it was English-Spanish. The investigators suggest that the order by sex interaction is attributed to rapport:

> ...the more highly Americanized boys responding more favorably to an initially English-speaking examiner, while the more restricted and less acculturized girls achieved better rapport with an initially Spanish-speaking examiner. (p.17)

In the second report, Anastasi and de Jesus (1953) administered the Goodenough Draw-a-Man Test to 50 Puerto Rican 5-year-old children. IQ scores were then compared with scores attained by similar groups of white and black children. No differences were found in IQ scores between groups. Analysis of the spontaneous language elicited from children showed that Spanish was used almost entirely during testing. Only about 2 percent of the words and less than 1 percent of the sentences were spoken in English. In addition, it was noted that Puerto Rican children exceeded both the white and black groups in mean sentence length and in maturity of sentence structure. These data are important because they illustrate that IQ and language development of children are not lower than other ethnic groups when attention is given to culture-fair methods of assessment.

Altus (1953) commented on the construction of the Wechsler tests of intelligence. She observed that discrepancies between verbal and performance IQ scores are expected to be small, regardless of size of score. To determine whether this was also true for Mexican children, WISC scores of 12 children referred for preliminary screening for special classes for the mentally retarded were studied. Mexican children attained a mean verbal IQ of 72.07 and a mean performance IQ of 84.01 whereas a non-Mexican comparison group showed a difference of only 2.5 ponts between the two scores in favor of the verbal scale. These data add further confirmation to the argument that the pattern of IQ scores for Hispanic children differ from non-Hispanic groups.

Another study pertaining to the performance of bilingual subjects on tests of intelligence administered in two languages is provided by Keston and Jimenez (1954). Form M of the 1937 Stanford-Binet was given in English to 50 fourth-grade school children. One month later, Form L was given in Spanish to the same subjects. The mean IQ of the Form M-English administration was 86.0 and on the Form L-Spanish it was 71.8. The expected correlation between Forms L and M of the Stanford-Binet is .93. In this study, the obtained correlation between the two forms was only .36. These findings are in direct contradiction to the earlier reports of Mitchell (1937) and Mahakian (1938), who found better performance on a Spanish version of the Otis Scale. The findings are also in opposition to Sanchez (1934) and Anastasi and Cordova (1953), who reported increased test scores on the second administration of a test of intelligence.

Three criticisms of the Keston and Jimenez study suggest the findings are tenuous and therefore of relatively little value for our purposes. First, the study is methodologically weak since Keston and Jimenez did not counterbalance the language order of their tests. The second and more important criticism is that the translation of Form L was provided by a professor of Spanish in Madrid, Spain. Thus, one must consider the possibility that the translation, even if technically accurate (and there is no reason to suspect it was not), probably omitted contemporary regionalisms, colloquialisms, and idoms. These children actually may have been tested in a third "foreign" language: Spanish from Spain. Finally, no accurate assessment of degree of bilingualism was attempted. Bilingualism was determined through a personal interview, a practice the authors justify with the incorrect assertion that "...no accurate measure of bilingualism is obtainable" (p. 264). The Hoffman Bilingual Schedule has been used widely by other investigators since 1934.

No other relevant studies were found from 1954 until 1960.

Developments in Testing Since 1960

In an extension of Ammons and Aguero's earlier study on the Full Range Picture Vocabulary test (FRPV), Norman and Mead (1960) note that most studies using Hispanic subjects limit themselves to the lower age groups and school grades because of the large dropout rate. To study subjects old enough to be in senior high school and college, Norman and Mead selected 150 volunteers for military services being processed through the Armed Forces Examining Station in Albuquerque, New Mexico. Fifty subjects from three age groups (17-18-19) were given the Hoffman Bilingual Schedule, followed by the FRPV.

Significant correlations were found between bilingualism and the FRPV (-.49) and between schooling and FRPV (.62). Among urban subjects, higher FRPV scores and less bilingualism were found. When obtained scores were compared with normative data for the FRPV, it was shown that Spanish Americans scored lower than Anglos. Upon combining these data with those of Ammons and Aguero, it was found that these differences between the two groups increase consistently from ages 7 to 19.

Another study using older subjects sought to assess the effects of speed on intelligence test performance. In this study, Knapp (1960) tested 100 Mexican and 100 Anglo males, mean ages 27.3 and 31.4, respectively, on the Cattell Culture Free Intelligence Test, Forms 2A and 2B. Subjects were administered both forms in a counterbalanced design; one form was given with the usual speed requirements and the other without time restrictions. It was hypothesized that the Mexicans, being less "geared to the clock," would perform differently on the two forms of the test.

As expected, all subjects achieved higher scores when permitted unlimited time (power) and lower scores when required to adhere to a time limit (speed). More importantly, the difference was significantly greater for Mexicans than for Americans. It was also found that the Mexican sample scored significantly lower than the American sample, which makes the "culture fairness" of the test suspect. Both groups significantly increased their scores from the first to the second administration, but an intergroup difference was not significant. Also, both groups evidenced higher scores when the order of testing was power-speed.

An innovative investigation on the intelligence of Hispanic children was conducted by Jensen (1961). In his introductory remarks Jensen notes that:

> Teachers of the mentally retarded have at times noted that on the playground the Mexican-American children appear to be normally bright as compared with the Anglo-American retarded children, even though all are educationally retarded. Thus one immediately suspects a basic difference underlying the intellectual handicap of the two groups. (p. 148)

Jensen then argues that a better way to determine whether a child is retarded is to give him standard learning tasks and to record latencies. To test this hypothesis, "bright" and "dull" Mexican-American and Anglo-American children were compared on a series of tasks, which included recall, serial learning, and paired associate learning. Results clearly showed a significant difference in the learning ability of bright and dull Anglo-American fourth- and sixth-grade children. A similar comparison between bright and dull Mexican-American children indicated that they did not differ in learning rate, despite a mean difference of 34 IQ points on the California Test on Mental Maturity. Comparing the two "dull" groups, Jensen found faster learning for the Mexican Americans.

These data confirm an implicit assumption being made here. IQ tests appear to be successful in discriminating between intellectual levels of Anglo Americans but not Hispanic groups. Jensen himself concludes:

> Considering the very large proportion of low IQs in the Mexican-American population from which this study drew its samples, the findings of the study are consistent with the hypothesis that the distribution of basic learning abilities in Mexican-Americans is not substantially different from that in the Anglo-American population of comparable socioeconomic level. (p. 156)

In sum, the findings reported by Jensen indicate that low IQ attainment of children does not necessarily imply mental retardation as defined by "slowness" in learning. IQ-test users must rely on additional measures to assess the true learning potential of these children.

Lesser, Fifer, and Clark (1965) examined the mental abilities of 6- and 7-year-old children of middle and low social class from four ethnic groups (Chinese, Jewish, Negro, and Puerto Rican). A modification of the Hunter College Aptitude Scales for Gifted Children was utilized. Although the scales were originally designed for use primarily with gifted children between the ages of 4 and 5 1/2 years, Lesser et al. reasoned that the scales would serve in measuring the abilities of older children. The mental abilities examined were verbal ability, reasoning, number facility, and space conceptualization. An analysis of covariance design was employed in order to control for effort and persistance, persuasibility or responsiveness to the tester, and age of the subject.

On the verbal ability scale, Jewish children ranked first (being significantly better than all other ethnic groups), Negroes ranked second and Chinese third (both being significantly better than Puerto Ricans), and Puerto Ricans fourth. On reasoning, the Chinese ranked first and Jews second (both being significantly better than Negroes and Puerto Ricans), Negroes third, and Puerto Ricans fourth. In terms of numerical ability, Jews ranked first and Chinese second (both being significantly better that Puerto Ricans and Negroes), Puerto Ricans third, and Negroes fourth. Finally, on space conceptualization, Chinese ranked first (being significantly better than Puerto Ricans and Negroes), Jews second, Puerto Ricans third, and Negroes fourth.

Of major importance here was the finding that ethnic group affiliation significantly affected the pattern of mental abilities. This is important since it indicates as Lesser et al. note "... that differential importance and emphasis is attached to these various mental skills in the different ethnic groups" (p. 72). Further, once the pattern specific to the ethnic group emerges, social-class variations within the ethnic group do not alter this basic organization.

These findings are of importance because they suggest that cultural influences in intellectual performance must be reassessed both in terms of measurement of intellectual capacity and educational planning for culturally different people. Cohen (1969), for instance, has shown that certain individuals and some cultural groups differ in conceptual styles. She has also shown that nonverbal tests of intelligence and achievement which have been designed to specifically reduce cultural bias remain biased against subjects who maintain a different conceptual style than that around which the test is oriented.

Reporting on the age-intelligence relationship of Puerto Rican adults, Green (1969) presents information drawn from the standardization sample for the Spanish language Wechsler Adult Intelligence Scale (WAIS). Four stratified, random groups, ages 25-29, 35-39, 45-49, and 55-64, are compared with similar age groups to test the prediction found in the literature that adult intelligence peaks at age 20 and declines thereafter.

Green's data show that WAIS full-scale scores rise to age 40 and never decline; verbal means rise to age 50 and never fall; performance scores, however; show a slight decline after age 40 -- due primarily to increased errors on "digit symbol." These data are consistent with corresponding investigations of the age-intelligence relationship of Anglo-Americans.

An interesting study by Christiansen and Livermore (1970) investigated the influence of social class on IQ scores. In this study, Christiansen and Livermore compared the performance of Anglo-American and Spanish-American school children from low and middle social classes on the verbal, performance, and full-scale scores on the WISC and on four WISC factors derived from Cohen's (1959) analysis of the WISC. These factors are freedom from distractibility, a measure of how effectively the subject attends and/or concentrates on a task; and relevance, a measure of the application of judgment to situations; the ability to use verbal skills in new situations -- a transfer factor; verbal comprehension, a measure of that aspect of verbal knowledge retained from former experiences; and perception organization, a measure of how effectively the subject can interpret and/or organize visually perceived materials.

The results indicated that middle-class children outperformed lower class subjects on all of the seven WISC scores and factors examined. Yet, Anglos scored higher on only four factors; full-scale IQ, verbal-scale IQ, verbal comprehension, and relevance. These findings clearly support the conclusion that the Hispanic children score lower on the WISC because of deficiency in English verbal fluency rather than innate intellectual capacity.

Another four investigations reported in the literature have used the WISC to assess the IQ's of Hispanic children. Simpson (1970) reports that correlations between WISC and WAIS scores are lower than expected among blacks, Anglos, and Mexican Americans who are below average intellectually. In this study it was shown that IQ scores are consistent for all three groups on the WAIS. In a study of examiner style, Thomas et al. (1971) show that the WISC scores of school-age Puerto Rican children are markedly affected by differences in examiners who were equivalent in sex, ethnicity, fluency in Spanish and English, and clinical experience. Specifically, Thomas et al. found that their subjects had higher WISC scores when the examiner encouraged active participation, verbalizations, and persistence on the child's part.

Killian (1971) administered the WISC, Illinois Test of Psycholinguistic Abilities, and the Bender Visual Motor Gestalt test to 84 subjects including Anglo-Americans, Spanish-American "monolinguals," and Spanish-American "bilinguals." Subjects were matched for school achievement, and bilingualism was determined by an interview with a native Spanish speaker. The three groups differed significantly on WISC scores, with the Anglo children scoring highest on all WISC scales and the Spanish-American bilinguals scoring the lowest. Supporting analyses from the other test instruments, however, suggest that the Spanish-American children were deficient on the input side of communicative skills, especially in understanding sentences and pictures. Killian's data also appear to suggest the bilingualism per se is not as important as earlier believed. He states:

> ...bilingualism...may not be as important in school performance as in the more complex combination of variables making up the ethnic class Spanish American. Poverty, restriction of experience, and different value systems are associated with bilingualism. (p. 43)

In an effort to assess the difference between WISC scores obtained by English and Spanish administrations of the test, Palmer and Gaffney (1972) tested bilingual children first in English and 1 year later in Spanish. No difference was found in IQ scores between the two test conditions. This finding stands in contradiction to those of Anastasi and Cordova (1953) and Keston and Jimenez (1954).

Hertzig and Birch (1971) use the Stanford-Binet Form L to test white middle class and Puerto Rican working-class children at 3 and at 6 years of age. The subjects were part of a group involved in longitudinal studies of behavioral development since the age of 3 months. In the case of the Puerto Rican children testing was conducted in either Spanish or English, depending upon the child's language of greatest fluency. Mean IQ scores tended to increase from age 3 to age 6 in both groups. These data therefore provide no evidence to support the hypothesis that measured intelligence declines with age among socially "disadvantaged" children. It should be noted, though, that the white middle-class children did score significantly higher on both testings. One interesting finding of this investigation was that the Puerto Rican children who were untestable at 3 years had IQ's at age 6 that were almost 13 points below those of the testable Puerto Rican children. This difference was highly significant statistically. Moreover, 50 percent of the initially untestable Puerto Rican children had IQ's below 90 at age 6, in contrast to 15 percent of the remaining Puerto Rican children who had IQ scores at this low level. Thus, untestability at 3 was a good predictor of poor performance at 6 for the Puerto Rican children but had no predictive value for their middle-class counterparts.

At this point some recent work by two investigators (Mayeske 1971 and Mercer 1971; 1972; 1973) working independently must be discussed in detail because their findings support a thesis implicit throughout this article -- that is, that tests of intelligence are "Anglo-oriented" and class-related. Mercer examined intelligence test scores, measures of adaptive behavior and sociocultural variables among Anglo, black, and Chicano subjects whose IQ scores were below average -- IQ score less than 85. Using the American Association on Mental Deficiency definition of mental retardation -- subaverage general intellectual functioning when associated with impairment in adaptive behavior -- Mercer found that when she compared subjects who had IQ's below 70 (the lowest 3 percent) Anglos consistently scored in the lowest 3 percent on the adaptive behavior scale. This was not true, however, of either blacks or Chicanos -- 91 percent of the blacks and 60 percent of the Chicanos with IQ's below 70 had passed the adaptive behavior test. In other words, intellectually subnormal Anglos required supervision while a significant proportion of the ethnic minority group members with equally low IQ scores were able to function in their environment without supervision.

The implication of Mercer's findings is that with regard to intelligence test scores, the tests are valid for Anglos, but not for blacks or Chicanos. Mercer (1973) concludes that:

> IQ tests now being used are, to a large extent, Anglocentric. They tend to measure the extent to which an individual's background is similar to that of the model cultural configuration of American society. Because a significant amount of the variance in IQ test scores is related to

> sociocultural characteristics, Mercer concludes that sociocultural factors must be taken into account in interpreting the meaning of any individual score.

In determining the influence of sociocultural factors, Mercer obtained information on the sociocultural characteristics of the families of her subjects. Mercer was then able to isolate the most significant sociocultural factors which influenced IQ. For her Chicano sample the five most significant sociocultural characteristics were:

- living in a household in which the head of household has a white-collar job;
- living in a family with five or fewer members;
- having a head of household with a skilled or higher occupation;
- living in a family in which the head of household was reared in an urban environment; and
- living in a family in which the head of household was reared in the United States.

Further analyses of these sociocultural characteristics with IQ scores revealed that the greater the number of these five characteristics that a Chicano subject possessed the higher was his IQ score likely to be. These findings are shown in Table 1.

Table 1 Shows that of those 25 subjects who possess all five of the identified sociocultural characteristics, a mean IQ of 104.4 is higher than the mean standardized IQ of 100 considered to be average for Anglo subjects.

The conclusion here is obvious:

> ...what the IQ test measures, to a significant extent, is the child's exposure to Anglo culture. The more "Anglicized" a non-Anglo child is, the better he does on the IQ test. (Mercer 1972, p.95)

Mayeske (1971), in a reanalysis of the achievement scores of sixth-grade students obtained by the Equality of Educational Opportunity study collected in the fall of 1965 at the direction of the Civil Rights Act, found that the difference in academic achievement performance between six racial ethnic groups (Indian, Mexican American, Puerto Rican, black, Asian, and white) is reduced when a variety of social-condition variables are corrected for statistically. The conditions of importance in the reanalysis were: social and economic well being of the family, the presence or absence of key family members, the student's and parents' aspirations for his schooling, their beliefs about how the person might benefit from an education, the activities that they engaged in to support these aspirations, one's region of residence, and the achievement and motivational levels of the students one goes to school with.

Table 1

Mean IQ scores for Chicano children identified as possessing different numbers of sociocultural characteristics related to IQ test performance.[1]

	Number of sociocultural characteristics				
	1-0	2	3	4	5
Number of cases	127	146	126	174	25
Mean IQ	84.5	88.1	89.0	95.5	104.4
Standard deviation	11.3	11.6	11.8	12.1	10.4

[1] Adapted from Mercer (1972).

Table 2 presents the average achievement scores for six racial-ethnic groups before and after statistical correction for the social condition variables. The achievement scores are composites of reading comprehension, mathematics achievement, and verbal and nonverbal ability. Inspection of Table 2 indicates that before correction (with a distribution mean of 50) whites attain the highest score with Asians following them by about 4 points. Approximately 5 to 7 points below them lie the Indians, Mexican Americans, and blacks, with the Puerto Ricans following these groups by another 4 points. Following correction for the social condition variables, we see that the variance in the achievement scores between the groups is considerably reduced.

Mayeske appropriately concludes from his analysis that "no inferences can be made about the 'independent effect' of membership in a particular racial-ethnic group on academic achievement" (p. 21) since achievement performance is almost completely confounded with numerous social conditions. These findings substantiate the results reported by Mercer (1971; 1972; 1973).

In concluding this part of the review of intelligence testing, we observe that since 1960 the quality of the research on intelligence has improved substantially. More sophisticated research designs have been employed and a wide variety of hypotheses have been tested. Nonetheless, the research lacks the coherence of systematic investigation. With the exception of Mercer's work, programs of research are not apparent from the literature review; no single variable or hypothesis was thoroughly examined by the numerous investigators who were working independently.

Table 2

Average composite achievement score classified by racial-ethnic group membership.[1]

Racial-ethnic group	Uncorrected mean achievement	Corrected mean achievement
Mexican American	42.2	47.5
Puerto Rican	38.6	47.1
American Indian	44.2	48.7
Black	42.5	49.4
Asian	49.4	51.2
White	53.2	50.4

[1] Adapted from Mayeske (1971).

Some Observations From the Review of Literature

There is now a body of literature on the testing of Hispanics that extends back slightly more than 50 years. Throughout this review a number of critical factors have emerged which are still central concerns in the testing of Hispanics today. Let us review these critical factors.

Language. Sheldon (1924) was the first to observe that children who understood English "imperfectly" performed below the level of "white children" who were monolingual speakers of English. This same observation was made by Paschal & Sullivan (1925) and Garretson (1928). Even as early as 1932, Sanchez on the basis of these data argued cogently that many Mexican American children would be at a disadvantage if tested in English. It was also Sanchez (1934a, b) who suggested that a simple translation of a standardized IQ test was not an appropriate solution to the testing of Spanish language dominant children. This point was further reinforced by the criticisms of the Keston & Jimenez (1954) study which employed a translated version of the Stanford-Binet which was possibly as difficult as the English version of the same test.

Now an extremely important consideration with respect to language is the question of degree of bilingualism. For instance, the subjects in the Keston & Jimenez study may have been Mexican American and familiar with Spanish, but not sufficiently bilingual enough to be tested in Spanish. A basic misconception concerning bilingualism is that because someone can speak in a second language they can also be tested in a second language. This is not always true. There are degrees of bilingualism and some children may come from homes where Spanish is the preferred language of the parents, but the children themselves have only a minimal expressive ability in Spanish. They are functionally receptive bilinguals only. That is, they understand Spanish and can follow their parents'' instructions, but when required to speak reply in English. Their parents in turn are also receptive bilinguals who understand English, but are inhibited in their expression in English. This language switching is quite common among second and third generation Hispanic families. Accordingly, an Hispanic child tested in Spanish may understand what is required or what is being asked, but may not be able to reply. Johnson's (1953) attempt to correlate degree of bilingualism and IQ test performance is the only study to date to investigate this important dimension. Killian (1971) does suggest that bilingualism, however, is not as important a factor as is socioeconomic status. This is a point that will be raised again, but it should suffice to say that the question of a language handicap and/or various degrees of bilingualism will occur most frequently with lower socioeconomic status Hispanics.

Performance Tests. Paschal & Sullivan (1925), more than 50 years ago, suggested that performance tests might be a fairer way to assess the ability of children who were handicapped when tested in an unfamiliar language. This suggestion was reinforced by Garth et al., (1936). A variant of this was also suggested by Anastasi & de Jesus (1953) who used the Goodenough Draw-A-Man test to assess the IQ of Puerto Rican children. It was also Anastasi & de Jesus who argued that culture-fair assessment included an examiner who could communicate with the children in both their languages and who understood their particular cultural circumstance. The only other study of examiner effects was that by Thomas et al (1971). Although Thomas et al. employed the WISC they did show that when the examiner was of the same sex and ethnicity they were more successful in coaxing Puerto Rican children to be more verbal, active, and persistent in their performance on the test. This served to increase the overall scores of the Puerto Rican children.

Knapp (1960) in an interesting study which has not been replicated found that Mexican Americans did significantly better on the Culture Free Intelligence Test when the speed requirement was eliminated and they were given unlimited time. As noted earlier in this review, these data call into question the "culture-fairness" of this instrument. Accordingly, it is possible that the task characteristics of a test such as speed are more important than whether a test is verbal, performance, or "culture-fair." It is not clear why this variable of speed should be important in the testing of Hispanics, but it clearly requires more examination than it has received to date.

Acculturation. Mitchell (1937) after noting that children performed better in Spanish than in English on the Otis Group Intelligence Scale suggested that a corrective score should be added to the score of a child if tested in English. This innovative idea fell upon deaf ears and was not implemented, but is reminiscent of a suggestion seen recently in the work of Mercer (1977). Before discussing Mercer in more detail, the term acculturation should be defined. For the sake of brevity we will here define acculturation as the acquisition of the language, customs, and social mores of one group by a member of another cultural group. This is an extremely important process in the life circumstance of many Hispanic Americans and a process that has largely been ignored in the testing literature. The process of acculturation has been central to our discussion here of bilingualism. Regardless of whether or not Hispanic parents want their children to become acculturated to the majority group, the process is occurring. The important factor is with the rate of acculturation. It appears that some Hispanics acculturate much more rapidly than other Hispanics. The parameters that control rate of acculturation have not been studied, and for our purposes here this is a dimension that cannot be ignored.

Altus (1949) was an early observer of the importance of acculturation in testing. He noted that even though Mexican Americans may have been born in the United States there were vestiges of the Mexican culture (e.g., language) which survived for several generations and which were important when evaluating the test performance of these individuals. A little later, Anastasi & Cordova (1953) suggested that Puerto Rican male children acculturated faster than their female counterparts. This then explained why their male subjects performed better in English and the females in Spanish when tested with both English and Spanish speaking examiners.

Mercer (1977) in her System of Multicultural Pluralistic Assessment (SOMPA) combines the WISC-R and th Adaptive Behavior Inventory for Children (ABIC) in arriving at an IQ score for Chicano children. Accoring to Mercer:

> The WISC-R is treated as a social system measure because it is designed to correlate with successful performance in the student role in the social system of the school. The ABIC is a social system measure because it measures the child's role behavior in the family, community, peer group, nonacademic school situations, self-maintenance, and earner-consumer roles form the viewpoint of the family. (p. 163)

The SOMPA is also based on four sociocultural scales. Mercer identifies these as: urban acculturation, socioeconomic status, family structure, and family size. Each of these scales carries with it a weighted score which is added to the score obtained on the WISC-R and which Mercer argues is more representative of the Chicano child's true intellectual potential. In effect this is a worked out version of the "correction" factor suggested by Mitchell (1937) some forty years earlier.

Mercer's Sociocultural Scales require further elaboration. Urban acculturation is composed of four factors: anglicization, sense of efficacy, community participation, and urbanization. Socioeconomic status is composed of two factors, occupation of head of household and source of income, as is family structure which is determined by marital status and relationship of child to parents. The Family Size Scale consists of only one factor which also identifies the scale. Most of the nine factors that make up the four Sociocultural Scales are in fact tied to degree of acculturation. Accordingly, a child who scores high on the four Sociocultural Scales is highly acculturated and possibly may not need the "corrective" score adjustment proposed. It is the unacculturated child who needs the adjustment score provided by Mercer's SOMPA method. As a matter of fact, Mercer suggests that her method is especially useful for identifying the gifted Chicano child who does not share the sociocultural characteristics of majority group children on which most tests have been standardized.

Acculturation and socioeconomic status are the two most critical factors involved in determining what use can be made of a test score obtained from an Hispanic. An Hispanic who is highly acculturated and who is not socioeconomically disadvantaged can be tested on any assessment instrument without any difficulty and the score interpreted as it would be for any individual who is a member of the majority group. The problem in testing is when the individual is not acculturated and also is socioeconomically disadvantaged. The test maker and consumer, if interested in the valid interpretation of test scores where Hispanics are concerned, have two alternatives: adopt a system such as the developed recently by Mercer or devise another method of assessing degree of acculturation that can be administered along with the test instrument. Unfortunately, although much has been said about acculturation, or lack of acculturation with Hispanics, there is really very little empirical work to guide test makers or users on what are the important dimensions involved in the acculturation process. Recently, Olmedo and Padilla (1978) have described a scale of acculturation that appears to have high empirical as well as construct validity. Much more empirical research on acculturation is necessary.

Before concluding this section a few more comments on acculturation are required. Hispanics constitute a subpopulation that is growing within the United States. As stated earlier, the 1970 census estimated the Hispanic population to be 11 million. Because of a higher birth rate and both legal and illegal immigration, the 1980 census may indicate as many as 20 million Hipanics residing in the United States. This figure does not include the 3.2 million inhabitants of the Commonwealth of Puerto Rico many of whom move freely between Puerto Rico and the United States. It is already estimated by census officials that Hispanics will constitute the largest minority group by the year 2000. These facts coupled with the emphasis being given to cultural maintenance within the Hispanic community are important since they underscore a final point that must be made here. That is, the the problems associated with bilingualism and acculturation will not disappear in the foreseeable future. It is not misleading to state that the problems associated with testing Hispanics discussed by Sanchez (1932a, b; 1934a, b) forty-five years ago are even more serious today.

Recommendations

Psychological assessment instruments can be classified into four broad categories: clinical, counseling, educational, and industrial. In this review we have been confined to educationally related instruments. However, the recommendations offered here should in large measure be applicable to all four categories of tests.

Culturally sensitive testing. This is the most elementary recommendation and was first suggested by Sanchez (1932a, b). A program of culturally sensitive testing should include: 1) the assessment of noncognitive characteristics (e.g., perceptual, psychomotor, affective, psychosocial, and behavioral), as well as cognitive abilities; 2) greater information sharing between the psychometrician, other school personnel, and the parents on the assessment instruments and outcomes in order to plan the best possible educational program for the chid; and 3) the employment of psychometricians who have a good understanding of Spanish and an appreciation of the important sociocultural variables that may affect the outcomes of the testing of Hispanic individuals.

Mercer's SOMPA procedure for testing Chicano children can clearly be placed within this category of culturally sensitive testing. The method acknowledges cultural as well as socioeconomic differences and proposes a method for adjusting test scores. The lead taken by Mercer, and incidentally published by the Psychological Corporation, should be expanded to include tests other than those falling within the category of educationally related instruments.

Criterion referenced tests. The call for criterion referenced tests has come from a wide variety of individuals involved in the use of tests (see for example, Houts, 1977). The value of criterion referenced tests are that they serve educational purposes more effectively by presenting information in terms of actual skills to be mastered by the individual without any comparisons with the performance of other individuals that could lead to unwarranted bias and discrimination (Olmedo, 1977). There are still some problems involved in criterion referenced measures, for instance, issues such as who determines the criteria and what is to be included in them must be thoroughly discussed if criterion referenced evaluation is to be used successfully (Olmedo, 1977).

A good example of these problems appeared when the author was asked to review the testing procedure which had been used by the personnel department of a major city to select a welfare department district supervisor for an area that was approximately 90 percent Hispanic. The entire test procedure was designed in a fashion to conform with criterion referencing and was in fact a very good examination. The problem, however, was that no consideration was given to the fact that the district supervisor of this particular area would have to work with and resolve problems for individuals who had no facility in English. Also a part of the district supervisor's responsibilities included meeting with community groups to explain welfare department policies. Both welfare department employees and community leaders stated that knowledge of Spanish and Hispanic culture were essential for a supervisor to carry out the responsibilities of the posiion. Yet these skills were not included on the criterion-referenced examination procedure used by the personnel department! Accordingly, a well intentioned attempt to design an examination procedure based on the job related tasks and skills essential for the position were unintentionally not included by the individuals who constructed the selection instrument. The test constructors, in this case, were not sensitive to the importance of language and culture and thus excluded these criteria from their instrument.

Criterion-referenced evaluation is most important in educational and industrial testing. The real impact of this form of evaluation is still to be felt. When serious efforts are made to design and construct criterion-referenced instruments attention should be given to language and cultural considerations of Hispanics.

Models of culture-fair assessment. Olmedo (1977) has reviewed the advantages and disadvantages of the regression, constant ratio and conditional probability models which have been proposed recently for making culture-fair selection decisions. These models differ in the magnitude of Type I and Type II errors they will tolerate in making decisions. Based on his review, Olmedo recommends that decision theoretic approaches be based on utility theory. Olmedo states:

> This approach has the advantage of providing a general model that requires the values or utilities in any given situation to be explicitly stated and entered into the decision-making process. If utilities are conflicting for the different parties involved, such as educational institutions and minority applicants, then this approach provides for explicity quantitative statements of the conflicting values. Thus, through bargaining, a compromise may be reached as to a set of utilities mutually acceptable to all. (p.191)

Several utility models have been proposed (Darlington, 1976; Petersen & Novick, 1976; Sawyer, Cole & Cole, 1976) recently. How adequate these models will be is still not clear, but they do appear promising and should be given greater consideration in future research. The value of this approach is that (1) such models provide for quantifying the degree of desirability associated with each outcome for the various parties affected by the decision process; and (2) the "expected utility" of each decision takes into account the probability of occurrence of the various outcomes in addition to their utility (Petersen & Navick, 1976).

References

Altus, Grace T. WISC patterns of a selective sample of bilingual school children. Journal of Genetic Psychology, 1953, 83(1), 241-248.

Altus, William D. Racial and bilingual group differences in predictability and in mean aptitude test scores in an Army Special Training Center. Psychological Bulletin, 1945, 42(5), 310-320.

Altus, William D. A note on group differences in intelligence and the type of test employed. *Journal of Consulting Psychology*, 1948, *12*(3), 194-195.

Altus, William D. The effect of adjustment patterns on the intercorrelation of intelligence subtest variables. *Journal of Social Psychology*, 1949, *30*(1), 39-48.

Anastasi, Anne, and Cordova, Fernando. Some effects of bi-lingualism upon the intelligence test performance of Puerto Rican children in New York City. *Journal of Educational Psychology*, 1953, *44*(1), 1-19.

Anastasi, Anne, and de Jesus, Cruz. Language development and nonverbal IQ of Puerto Rican preschool children in New York City. *Journal of Abnormal and Social Psychology*, 1953, *48*(3), 357-366.

Ammons, Robert B., and Aguero, Abelardo. The Full-Range Picture Vocabulary Test: VII. Results for a Spanish-American school-age population. *Journal of Social Psychology*, 1950, *32*(1), 3-10.

Carlson, Hilding B., and Henderson, Norman. The intelligence of American children of Mexicaqn parentage. *Journal of Abnormal and Social Psychology*, 1950, *45*, 544-551.

Christiansen, Ted, and Livermore, Gary. A comparison of Anglo American and Spanish American children on the WISC. *Journal of Social Psychology*, 1970, *81*(1), 9-14.

Cohen, J. The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *Journal of Consulting Psychology*, 1959, *23*, 285-299.

Cohen, R.A. Conceptual styles, culture conflict, and nonverbal tests of intelligence. *American Anthropologist*, 1969, *71*(5), 828-856.

Cook, John M., and Arthur, George. Intelligence rating of 97 Mexican-American children in St. Paul, Minnesota. *Journal of Exceptional Children*, 1951, *18*(1), 14-15+.

Darcy, Natalie T. The performance of bi-lingual Puerto Rican children on verbal and non-verbal language tests of intelligence. *Journal of Educational Research*, 1952, *45*(7), 499-506.

Davenport, E.L. The intelligence quotients of Mexican and non-Mexican siblings, 1932, *36*(923), 304-306.

Garcia, J. Intelligence testing: Quotients, quotas, and quackery. In J. Martinez (Ed.), *Chicano Psychology*. New York: Academic Press, 1977.

Garretson, O.K. A study of causes of retardation among Mexican children in a small public school system in Arizona. *Journal of Educational Psychology*, 1928, *19*(1), 31-40.

Garth, Thomas R. A comparison of the intelligence of Mexican and mixed and full blood Indian children. *Psychological Review*, 1923, *30*(5), 388-401.

Garth, Thomas R. The intelligence of Mexican school children. *School and Society*, 1928, *27*(705), 791-794.

Garth, Thomas R.; Elson, Thomas H.; and Morton, Margaret M. The administration of non-language intelligence tests to Mexicans. *Journal of Abnormal and Social Psychology*, 1936, *31*(1), 53-58.

Garth, Thomas R., and Johnson, Harper D. The intelligence and achievement of Mexican children in the United States. *Journal of Abnormal and Social Psychology*, 1934, *29*(2), 222-229.

Green, Russel F. Desarrollo y estandarizacion de una escala individual de inteligencia para adultos en Espanol. [The development and standardization of an individual intelligence scale for adults in Spanish.] *Revista Mexicana de Psicologia*, 1964, *1*(3), 231-244.

Green, Russel F. Age-intelligence relationship between ages sixteen and sixty-four: A rising trend. *Developmental Psychology*, 1969, *1*(5), 618-627.

Herr, Selma E. The effect of pre-first-grade training upon reading readiness and reading achievement among Spanish American children. *Journal of Educational Psychology*, 1946, *37*(2), 87-102.

Hertzig, Margaret, E., and Birch, Herbert G. Longitudinal course of measured intelligence in preschool children of different social and ethnic backgrounds. *American Journal of Orthopsychiatry*, 1971, *47*(3), 416-425.

Hoffman, B. *The tyranny of testing*. New York: Collier Books, 1964.

Houts, P.L. Editorial: Standardized Testing in America, II. *National Elementary Principal*, 1975, *54*(6), 2-3.

Houts, P.L. *IQ: The Myth of measurement*. New York: Hart Publishing Co., 1977.

Jensen, Arthur R. Learning abilities in Mexican American and Anglo American children. *California Journal of Educational Research*, 1961, *12*(4), 147-159.

Johnson, Granville B. Bilingualism as measured by a reaction-time technique and the relationship between a language and non-language intelligence quotient. *Journal of Genetic Psychology*, 1953, *82*(1), 3-9.

Keston, Morton J., and Jimenez, Carmina. A study of the performance on English and Spanish editions of the Stanford-Binet intelligence test by Spanish-American children. *Journal of Genetic Psychology*, 1954, *85*(2), 263-269.

Killian, L.R. WISC, Illinois Test of Psycholinguistic Abilities, and Bender Visual-Motor Gestalt Test performance of Spanish-American kindergarten and first-grade school children. *Journal of Consulting and Clinical Psychology*, 1971, *37*(1), 38-43.

Knapp, Robert R. The effects of time limits on the intelligence test performance of Mexican and American subjects. *Journal of Educational Psychology*, 1960, *51*(1), 14-20.

Kohn, S.D. The numbers game: How the testing industry operates. *National Elementary Principal*, 1975, *54*(6), 11-23.

Lesser, Gerald S.; Fifer, Gordon; and Clark, Donald H. Mental abilities for children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development*, 1965, *30*(7), 1-115.

Mahakian, Charles. Measuring the intelligence and reading capacity of Spanish-speaking children. *Elementary School Journal*, 1939, *39*(10), 760-768.

Manual, H.T., and Hughes, Lois S. The intelligence and drawing ability of young Mexican children. *Journal of Applied Psychology*, 1932, *16*(4), 382-387.

Mayeske, George W. "On the Explanation of Racial-Ethnic Group Differences in Achievement Test Scores." Unpublished manuscript, 1971. 29 pp. (Xerox).

Mercer, J.R. "Pluralistic Diagnosis in the Evaluation of Black and Chicano Children: A Procedure for Taking Sociocultural Variables Into Account in Clinical Assessment." Paper read at the Meeting of the American Psychological Association, Washington, D.C., September 3-7, 1971.

Mercer, Jane R. IQ: The lethal label. *Psychology Today*, 1972, 6(4), 44; 46-47, 95-96.

Mercer, J. *Labeling the mentally retarded*. Berkeley, California: University of California, 1973.

Mercer, J.R. Identifying the gifted Chicano child. In J. Martinez (Ed.), *Chicano Psychology*. New York: Academic Press, 1977.

Mitchell, A.J. The effect of bilingualism in the measurement of intelligence. *Elementary School Journal*, 1937, 38(1), 29-37.

Norman, Ralph D., and Mead, Donald F. Spanish-American bilingualism and the Ammons Full Range Picture Vocabulary Test. *Journal of Social Psychology*, 1960, 51(2), 319-330.

Olmedo, E.L. Psychological testing and the Chicano: As reassessment. In J. Martinez (Ed.), *Chicano Psychology*. New York: Academic Press, 1977.

Olmedo, E.L., & Padilla, A.M. An empirical and construct validation of a scale of acculturation for Mexican Americans. *Journal of Social Psychology*, 1978, 105, 781-790.

Palmer, Michael, and Gaffney, Philip D. Effects of administration of the WISC in Spanish and English and relationship of social class to performance. *Psychology in the Schools*, 1972, 9(1), 61-63.

Pasamanick, Benjamin. The intelligence of American children of Mexican parentage: A discussion of uncontrolled variables. *Journal of Abnormal and Social Psychology*, 1951, 46(4), 598-602.

Paschal, Franklin D., and Sullivan, Louis R. Racial differences in the mental and physical development of Mexican children. *Comparative Psychology Monographs*, 1925, 3(1), 1-76.

Sanchez, George I. Group differences in Spanish-speaking children: A critical review. *Journal of Applied Psychology*, 1932a, 16(5), 549-558.

Sanchez, George I. Scores of Spanish-speaking children on repeated tests. *Journal of Genetic Psychology*, 1932b, 40(1), 223-231.

Sanchez, George I. Bilingualism and mental measures. *Journal of Applied Psychology*, 1934a, 18(6), 765-772.

Sanchez, George I. The implications of a basal vocabulary to the measurement of the abilities of bilingual children. *Journal of Social Psychology*, 1934b, 5(3), 395-402.

Sheldon, William H. The intelligence of Mexican children. *School and Society*, 1924, 19(475), 139-142.

Shotwell, Anna M. Arthur performance ratings of Mexican and American high-grade mental defectives. *American Journal of Mental Deficiency*, 1945, 49(4), 445-449.

Simpson, Robert L. Study of the comparability of the WISC and the WAIS. *Journal of Consulting and Clinical Psychology*, 1970, 34(2), 156-158.

Thomas, Alexander; Hertzig, Margaret W.; Dryman, Irving; and Fernandez, Paulina. Examiner effect in IQ testing of Puerto Rican working-class children. *American Journal of Orthopsychiatry*, 1971, 41(5), 809-821.

Darlington, R.B. A defense of "rational" personnel selection, and two new methods. *Journal of Educational Measurement*, 1976, 13, 43-52.

Petersen, N.S., & Novick, M.R. An evaluation of some models of culture-fair selection. *Journal of Educational Measurement*, 1976, 13, 3-29.

Sawyer, R.L., Cole, N.S., & Cole, J.W.L. Utilities and the issue of fairness in a decision theoretic model of selection. *Journal of Educational Measurement*, 1976, 13, 58-76.

CASE STUDY
INSTRUCTION AND TESTING
SCHOOL DISTRICT OF PHILADELPHIA

S. Charp
Director, Division of Instructional Systems
School District of Philadelphia

Abstract

It has been documented that computer based instructional programs have made significant contributions to the educational process. In the Philadelphia School System, where computers have become an integral part of the curriculum since 1968, their impact has been substantial, with a variety of progams being initiated. However, the use of computers for testing purposes has not been very innovative, and there has been little experimentation. This has been due primarily to the limitations of the computer systems which have forced the use of true or false, multiple choice, or limited character response questions.

In the Computer Assisted Instruction programs, testing is done on-line, the results of these tests the basis for student activity. Tests are interspersed throughout the tutorial material as pre-tests, criterion tests, chapter tests, or multi-topic tests. In computer managed instruction, test results are used to determine student status and in the assignment of material. The assignment of tests on various difficult levels depends on the number of tests available and the sophistication of the program. The tests are based upon behavioral objectives and the test questions used are to check on mastery. The behavioral objectives not mastered are identified and a remedial prescription is presented. There is the pattern of pre-test, diagnosis, prescription, post-test, and reporting. This common and quite effective model is being applied in many areas.

Another type of system adapts instructional treatments and individual differences in students to available curricular options. This system is so designed to ensure that each student masters the basic minimum essentials in ways which conform to his individual learning characteristics. The program consists of an evaluation system which measures each child's learning characteristics and his mastery of stated objectives: a curriculum bank of independent learning packets are sequenced and coded according to objectives and various learning characteristics (including pre- and post-tests) before a computer management system which matches the learning activity to the needs and characteristics of the student. The measured learning charcteristics include reading level, ability level, learning style (auditory or visual) and cognitive style (abstract or concrete). The amount of testing is extensive, and results have shown students make significant gains. Each student is expected to meet the minimum mastery level (80%) of each objective he addresses. The locally developed material has been carefully constructed and validated. The student begins by selecting one of the highest priority objectives in which he has shown a deficiency. He is assigned a diagnostic test which determines his first "prescription"--the material which best matches his

learning characteristics and which may be the first in a sequence leading to mastery of the objective. A progress test to show mastery influences the "best match" packet on the next sequential step. In all of the programs, the computer tests students, keeps records, and provides prescriptions.

Many equivalent and non-identical test questions have been prepared to assist in student mastery. Tailored testing, which will necessitate a careful and an appropriate selection of items for ability banks, is bound to lead to improved measurement.

INTRODUCTION

Computers have been used for instruction by students in the Philadelphia Public Schools since 1968, particularly to meet the needs of low income inner city youth in augmenting their basic skill instruction in Mathematics and Reading Comprehension and Decoding. In addition to using the computer for enforcing basic skills, the computer is also being used in the areas of high school mathematics, science, social science, and business and vocational education. In computer assisted instruction, tutorial material is presented; in computer managed instruction, the computer tracks student performance and prescribes individual assignments based on criterion referenced testing. This paper will describe how testing is being handled in several computer based instruction systems that have been developed by the School District and are now being used in classroom environments.

COMPUTER ASSISTED INSTRUCTION (CAI)

In CAI, testing is part of the presentation. The courseware is divided into a modularized hierarchy of instructional levels. Each level utilizes a complete range of instruction strategies and techniques which allow full individualization. At each level of instruction the student is offered a pre-test, a CAI tutorial, guided practice review help sequence, and a mastery test. Learners are assessed for placement and diagnostic purposes by a series of inventory assessment instruments at each level. Progress, retention, and achievement are measured by mastery and retention tests.

Tests are used not only for ranking students and generating marks but also for making a decision--the decision on what each student should do next. Some students go on to more advanced material, others are given more explanation, more practice, or other forms of instructional remediation. Following remediation, the design calls for retesting to assure readiness for advanced learning. If readiness is not found, and the particular remedial activities are not satisfactory, the student is told to see his teacher. By careful analysis of the status of the student before remediation, the type of remediation that is offered and the learning outcomes, the teacher can determine how to improve student performance and also assess the remedial instruction portion of the program. Tests are generated randomly during question periods from stratified question banks. Specific lessons and tests for individuals are retrieved according to the student's needs. The computer automatically scores multiple choice and numerical tests.

ILLUSTRATION - CAI MATHEMATICS

Review of the instructional material within a chapter and chapter tests are administered on-line at end of each chapter. If a student fails to meet criterion on an outquiz (criterion quiz for an instructional block) a decision table is presented which gives the student the option of reviewing any part of the instructional material (instruction, practice, summary) covered by the quiz, or to sign-off. If the student takes the option to sign-off, he will begin at the decision table when he signs on again.

On-line tests are:

1. Pre-skills tests--a preliminary test for one or more instructional blocks
2. Pre-test for one or more instructional blocks
3. Criterion test at the completion of an instructional block
4. Chapter test--percentage correct in a chapter test

In the chapter tests, the test items parallel the format and content of questions presented in the instructional portion of the program and the on-line quizzes. They are viewed as criterion tests for the chapter. If a student's performance is unsatisfactory the area of difficulty is worked on. Each chapter test consists of a series of pools of test items. Test items are representative questions from the various blocks in a chapter. Items from one or more blocks are stored in a pool. Not all items in a pool are presented to the student. For example, there may be five test items from the first block of a chapter in an item pool, but the student receives only three from the pool. These are presented in random order, so that the probability of the same test item being presented in the same sequence by the program is greatly reduced. The student is given no knowledge of results during the execution of the test. After he has completed the test, he is given his score and may be referred to review questions. Each student gets the same review questions in the same order. Unlike the chapter-test questions, the student is given the correct answers if he answered incorrectly. A pretest copy of the Review Chapter Test is available which enables areas of difficulty to be identified and review materials to be assigned.

Following each instructional unit the students take a test to receive immediate feedback on their performance on each question. A multi-topic test is given after a certain number of units. This test consists of questions from those previously used, and are four-choice multiple choice questions.

COMPUTER MANAGED INSTRUCTION

Course material is divided into segments or instructional units. At the begining of each unit, the student is given a pre-test. This is administered by either the teacher or the computer. Where a collection of test questions is stored in the computer, the computer prints out copies for the student. The purpose of the pre-test is to determine the student's initial status in order to direct him to a specific learning task. The computer assigns learning tasks automatically by examining test scores. The subject matter is presented through reading assignments, programmed tests, CAI programs, audio-visual material, classroom instruction, etc. Throughout the material the student is given progress tests to determine whether he is progressing satisfactorily. These tests "prescribe" material on various difficult levels based on student performance. When a student has completed the assigned tasks, he takes a post-test covering the unit. The teacher receives various reports at the completion of each unit in order to follow the pattern of accomplishment of each of the students to determine where additional help is required. There is then, a basic pattern of pre-test, diagnosis, prescription, post-test and reporting. The computer has stored teacher-prepared lessons, tests, and decision algorithms.

The system reviews performance data on tests administered at periodic intervals through the learning sequence in order to reach a decision about the next learning assignment. The prescription is based on table look-up procedures which involve dividing the test score distribution into weighted scores and prescribing different learning assignments depending on where a particular score falls.

GENERAL COMPUTER MANAGED INSTRUCTION FORMAT

Course Organization

The course is based on behavioral objectives (B.O.s) pre-determined by authors as essential to the mastery of the subject. The course is divided into modules (Topics) which are subdivided:

- Pre-test based upon major B.O.s of the module to determine entering competencies

- A series of informational units in multi-media materials grouped in 3 or more levels of difficulty based on unit B.O.s. (In order to meet the student's needs in level of difficulty and type of learning media to facilitate mastery of B.O.s.)

- Unit Tests on 3 or more levels of difficulty based on unit B.O.s. (In order to determine mastery of the unit B.O.s as measured by the tests.)

- Post Tests based upon major B.O.s of the module to determine behavioral change and areas which need more study.

The Computer's Role

- Evaluates student success in mastering B.O.s as shown by his responses in B.O. based tests entered into computer
- Keeps records of individual progress of each student
- Manages the student's learning procedures by prescribing study assignments and determining level of difficulty based on past performance
- Prints reports for teacher as requested
- Prints analysis of course for research evaluation and modification.

Student-Computer-Teacher Interaction

A modular pre-test is taken, off-line, by the student. On completion, the student signs in on the terminal--his name, identification (ID) and the code word "test" to indicate to the computer that he wants his test evaluated. The computer identifies him, and, from its memory bank, verifies the test he should be taking. The student puts in each answer, which is evaluated. The score is printed and is also stored for future reports.

The teacher has the option of skipping a student based on the results of the pre-test analysis, by instituting a "skip" command, followed by the student's ID and a search for a new assignment. The skip can be to another modular pre-test, skipping over all units in the current module if the teacher so determines from the pre-test results, or the skip can be to any unit within the current module if the teacher so decides from the pre-test results. A skip is not allowed over a pre-test into another module since this would inhibit any measurement of behavioral change via pre-post test analysis.

In normal sequence the assignment for the first unit is printed out (if the assignment must be restated for some reason later, the student can then sign in his ID, type the code word "work" and the assignment will be reprinted for him). The assignment is for study off-line using multimedia material which is necessary for mastery of the B.O.s of the unit, and at the level of difficulty determined by a record of previous success.

When the student feels he has mastered the assignment, he gets, from the teacher, an off-line test for that unit at the indicated level. Upon completion of the test, the student signs in as before and types code word "test". Verification proceeds as before including the test number and level in stored memory. The student enters his answers, and the answers are evaluated in terms of the stored B.O.s for the unit. If the passing criterion is met, the student receives appropriate feedback, and a new assignment for off-line materials is made to help master the B.O.s for the next unit. The teacher may override this with a skip request to any other unit in the module. An exception to this procedure occurs when a student fails all questions for a particular B.O. In that case, even if he has met the criterion test score, he is assigned, after the appropriate explanatory statement, a prescription for that B.O., followed by a retest covering all the B.O.s of the unit. If evaluation of this meets the criterion, he is sent on.

If the criterion is not met, the behavioral objectives not mastered are identified and a remedial prescription for off-line work leading to mastery of those failed B.O.s is printed. He is not required to study material for B.O.s he has mastered. The remedial material is not only for the primary B.O. of a failed question, but for secondary B.O.s (sometimes for earlier units) whose mastery is faulty as indicated by the combinations of wrong choices made by the student in the test. The computer prints a prescription for off-line study of materials for the unique combination of B.O.s needed as shown by the student's response combinations. This prescription is for materials at a less difficult level of expression in varied media. Depending upon stored criteria, as it is matched against the test results, the student may be assigned work at the next less difficult level or on a still lower level.

After off-line study, the student is given an off-line test at the lower level of difficulty indicated. The sign-in-verification-evaluation procedure goes on as before. If the success criterion is met, the student gets a new assignment for the next unit, its level of difficulty based upon evaluation of his past levels of achievement.

If the criterion is not met, the B.O.s not mastered are recognized and another unique off-line assignment is printed, based upon his error combination. The cycle now repeats at a lower difficulty level. If the criterion of the next lower level test is met, he is sent on at an entering level indicated as optimal by his past stored achievement record.

The prescription-text-evaluation-represcription is repeated until the computer finds no more levels in its memory. If all levels are exhausted, a message is printed for the teacher informing him of this. The teacher then gives individual help in attaining mastery of the B.O.s needed as indicated by the last test. When the teacher is satisfied that mastery has been attained, he then gives the lowest level off-line test again. After completion, the student enters answers for evaluation as before. If he passes, he goes to the next unit at the proper level as described previously. If not, the teacher-test-evaluation cycle is repeated until success is achieved, or until the teacher manually overrides the computer evaluation. All results of all unit tests are stored for future requests for reports and analyses.

The teacher, at the end of any unit successfully completed, has the option of skipping a student back to any unit in the module for re-study, or of moving him forward within the module if he thinks some material can be skipped. The computer will check to see if the skip is not out of the module (past the post-test) and implements it with an assignment.

After all units in module have been passed as shown by the unit tests at some level, the post-test to measure mastery of the major B.O.s in the module is scheduled by the computer. This is taken off-line, and the answers typed in following the same general procedure as previously described. The test report is printed for student and teacher.

The teacher now has the option of requesting the computer to move the student back to any unit(s) in the module just completed, for further study, if, on the evidence of the post-test, he feels mastery is insufficient in some areas. He can request a skip at this time to the pre-test of any other module, forward or back in the course, if, in his judgment, it would be beneficial for the student to be so moved. Otherwise the student goes to next pre-test.

Reports and Record Keeping

Reports can be requested any time by a teacher. A status report prints the student's ID, his current module, unit, and level, the last unit and level passed, his current assignment (B.O.s to master), and the number of days he has spent in the unit. Also the date of the report is printed.

An individual student progress report, which is printed on demand by the teacher, lists the student's ID, and a cumulative record of his progress; including modules, units, levels, and test scores for every pre- and post-test and all unit tests passed or failed to date, together with the date of the paper.

Item analysis reports are used for future revision of the curriculum. A record is available of the cumulative pass-fail records for each B.O. as accumulated from student performance records.

CMI ELECTRONICS FORMAT

Successful completion of last three unit tests are examined by the computer:

- o If average is above set criterion, level "A" is assigned
- o If below criterion, level "B" is assigned. The student is referred to a worksheet which lists particular pages in tests, lists exercises, or other material needed to achieve behavioral objectives. Off-line tests are given to determine student's mastery. Test answers are entered into the computer for evaluation. If criterion (set for 70%) for "A" test is met, and even if only one question on any behavioral object was answered correctly, he may go on. If

both questions for the same behavioral objective are not answered correctly, the student is given an assignment for that B.O. If upper criterion is not reached and the score is above 40%, the computer checks to see if a "B" level assignment has already been given. If below 40% "C" level remediation and test are assigned.

o Remediation at "B" level is assigned, if not previously used, based upon student's unique combination of errors. "B" level test is entered and if criterion is met student goes on, if ""B" test criterion is not reached "C" remediation and test are assigned.

o If "B" level assignment has already been given, computer checks to see if "C" level assignment has been made. Assignment on "C" level is usually to various programmed booklets which have been developed (one for each B.O. for which remediation has been indicated.)

o If test "C" criterion is not met, teacher offers individual assistance and test "C" is given again. If test criteria "C" is met (70%) student will get assignment. If not, the following "C" cycle is repeated until "C" criterion test is met. (Choices of certain distractors may trigger additional remedial assignments to B.O.s of earlier units.) See Chart I.

GENERAL OUTLINE OF CMI/CAI STRATEGY

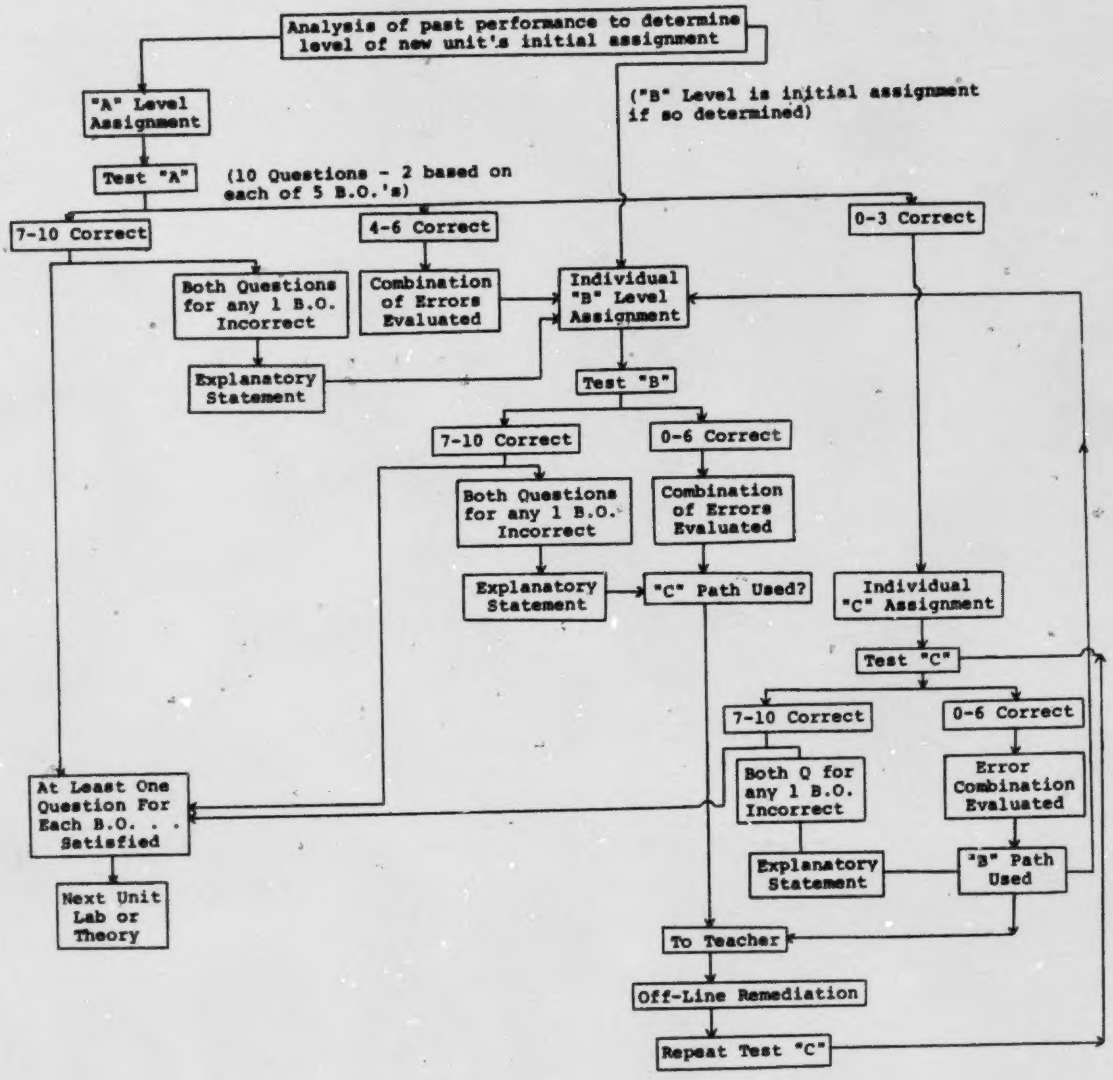
DESIGNER:

Determines block of content to be taught
Writes general objectives
Writes pre-test
Writes Criterion Test--one or more questions for each behavioral objective
Develops media and levels of presenting the material. Uses a variety of resources--text assignments, prepared lesson sheets, CAI material when available, prepared programmed material audio visual material, etc. All content material is identified or written for various levels of difficulty and geared to behavioral objectives. Programmed instruction may be written especially for lower levels of achievement. Enrichment material is provided for able students.

STUDENT:

Reads information designed to help reach behavioral objectives.
Takes criterion test on information. This test may be on-line or student may type in the answers only.
Uses a variety of resources to meet objectives.
Must pass a test at any one of three three levels in each unit in order to proceed to the next unit.
Can refer to "his file" to check on his own progress at any time.

CHART I: FLOW CHART - CMI


Analysis of past performance to determine level of new unit's initial assignment
"A" Level Assignment
("B" Level is initial assignment if so determined)
Test "A"
(10 Questions - 2 based on each of 5 B.O.'s)
7-10 Correct
4-6 Correct
0-3 Correct
Both Questions for any 1 B.O. Incorrect
Combination of Errors Evaluated
Individual "B" Level Assignment
Explanatory Statement
Test "B"
7-10 Correct
0-6 Correct
Both Questions for any 1 B.O. Incorrect
Combination of Errors Evaluated
Explanatory Statement
"C" Path Used?
Individual "C" Assignment
Test "C"
7-10 Correct
0-6 Correct
At Least One Question For Each B.O. . . Satisfied
Both Q for any 1 B.O. Incorrect
Error Combination Evaluated
Explanatory Statement
"B" Path Used
Next Unit Lab or Theory
To Teacher
Off-Line Remediation
Repeat Test "C"

COMPUTER:

- Analyzes answers for each behavior objective or combination of objectives.
- Sends student to individualized material for remediation or enrichment or new assignment (CAI Drill & Practice, audio visual programmed material).
- Keeps record of test results, individual and group progess.
- Prints out a variety of reports.

TEACHER:

- Works with individual student or with groups of students.
- Has immediate knowledge of the progress and needs of each student in his class.
- Can determine any variation in assignment--has final option on student progress.

AN INDIVIDUALIZED COMPUTER BASED INSTRUCTION MANAGEMENT SYSTEM BASED ON LEARNING CHARACTERISTICS

OVERVIEW

This system focuses exclusively on basic minimum objectives in various subject matter fields rather than encompassing a complete curricula in mathematics, science, communication arts, or social studies.

System is designed for maximum flexibility in responding to unique sets of objectives developed in the Philadelphia School System.

An effort is made to assist students in attaining a minimum standard of maturity in a variety of affective as well as the more traditional cognitive objectives.

The system attempts to adapt instructional treatments to individual differences in children.

The system consists of three basic components: a set of instruments and techniques for assessing student needs, a bank of curriculum packets related to assessed needs, and a computer based system for relating individual needs to availble curriculum options.

Information about the educational objectives and listings of additional enrichment experiences have been placed in the computer memory. When a student enters the system, the computer is given information concerning his previous educational experiences, his aptitudes, reading level, and other personal characteristics. The student's progress is assessed in terms of educational objectives which are to be mastered rather than by comparing the performance against that of some normative group.

STUDENT EVALUATION SYSTEM

Individual treatment is designed to determine (1) the child's instructional needs, (2) the most effective way to teach the child, and (3) the child's success in learning the materials. Four variables are measured in this process:

1. Reading Level--Standardized test results are used to indicate a student's reading level. The test results are categorized in one of three levels (0-2nd year, 3-5 year, 6 and above) and entered in the student's evaluation form.

2. Aptitude Level--Scores from mental ability tests are categorized as "below average" or "average" or "above average" for input into the student file.

3. Learning Style--Three distinct sensory pathways--the visual, auditory, and kinesthetic are used. "Learning style" is defined here as the profile of the student's relative ability to use and remember information processed through the three modalities.

Some students appear to have strengths or weaknesses in learning by means of one or another of these pathways. Three measures were constructed and administered to all students in order to assess their relative abilities in learning comparable tasks presented through each modality. A specific learning style is entered in the evaluation file only if a student shows marked superiority in one of these tests. Otherwise, an identification of a "mixed" code is used.

4. Cognitive Style--A student's mental processes may be especially suited to learning either abstract or concrete relationships. For example, one student might find that the learning of number systems is facilitiated by a considerable amount of practice with numbers, while another student might need to be taught only the principles involved. Again, a student is given a particular designation only if he shows large score differences on the measures of cognitive style.

CURRICULUM SYSTEM

This system focuses exclusively on basic school objectives which are defined as the minimal outcomes of a middle school educational experience (grades 5-8). The facts, principles, and operations that are expected of students at the completion of four years of instruction in mathematics, science, social studies, and communication arts have been identified. Added to such content objectives are a series of "skill" objectives defining the minimum levels of proficiency in more general skills which cut across the various subject matter areas. One example, under the general area of "Learning Strategy" is "independence in learning" defined as the ability to take an assignment requiring the use of reference materials, locate the designated materials in the library, obtain the required information and report it in a reasonably coherent manner. A number of affective as well as cognitive objectives are included in the "skill" category, such as "tolerance of individual differences" and "empathy toward others". The system concerns itself with 38 content and 29 skill objectives which are reviewed each year in light of the previous year's experiences. Tests were written for all content objectives and rating scales (to be used by both the students and teachers) were developed for evaluating students with reference to skill objectives.

Curriculum materials were prepared in order to provide effective learning treatments for students with specified learning aptitudes. All curriculum materials are presented in the form of learning packets, which are brief curriculum units involving independent or semi-dependent learning activities. Each packet attempts to teach one segment of a particular learning objective. For example, the concept of place value in numbers is one component of the objective "Arithmetic Skills". Enough variations are available for the various topics to provide for the variables that have been determined useful in structuring the learning experiences of children.

DIAGNOSTIC TESTS

Diagnostic tests are used to help determine to which of the available learning materials the student should be assigned. Mastery of most of the basic objectives requires the student to perform or comprehend a number of component operations or concepts, each representing one segment of the total objective. For example, the concept of place values in numbers is one component of Arithmetic Skills. The diagnostic test attempts to determine which segments of the objective have already been mastered by the student and should therefore be eliminated from his learning prescription.

PROGRESS TESTS

These tests are designed to measure mastery of the material presented in each learning unit. A variety of testing techniques are used, some on-line and some off-line. These include objective items, written response items, and discussion with the student. Whatever the testing technique used, a "pass-fail" decision is always reached for every packet the student completes. Results become part of the file.

PACKET CODING SYSTEM

A packet is defined as a curriculum unit involving independent or semi-dependent learning activities. Each packet consists of a statement of the packet objectives, a unit of instructional material and a progress test. Each packet is coded in terms of category (content or skill objective) area, (Communication Arts, Creative Thinking, etc.) objective, segment (specific concept or topic which is taught and the relative order of its presentation in the instructional sequence, reading level, difficulty level, learning style and cognitive style).

The structure for coding the treatment characteristics of packets for entry into the curriculum bank is shown in Chart II.

CHART II

CODED CHARACTERISTICS	CODING CATEGORIES	DATA BANK INPUT
Reading Level	0 - 2 grade	0
	3 - 5 grade	1
	6 and above	2
Difficulty Level	Simple	0
	Average or Above	1
Teaching Mode	Visual	0
	Auditory	1
	Kinesthetic	2
	Mixed	3
Cognitive Mode	Abstract	0
	Concrete	1
	Mixed	2

COMPUTER'S FUNCTIONS

The computer's function in managing the instructional program can be divided into three phases:

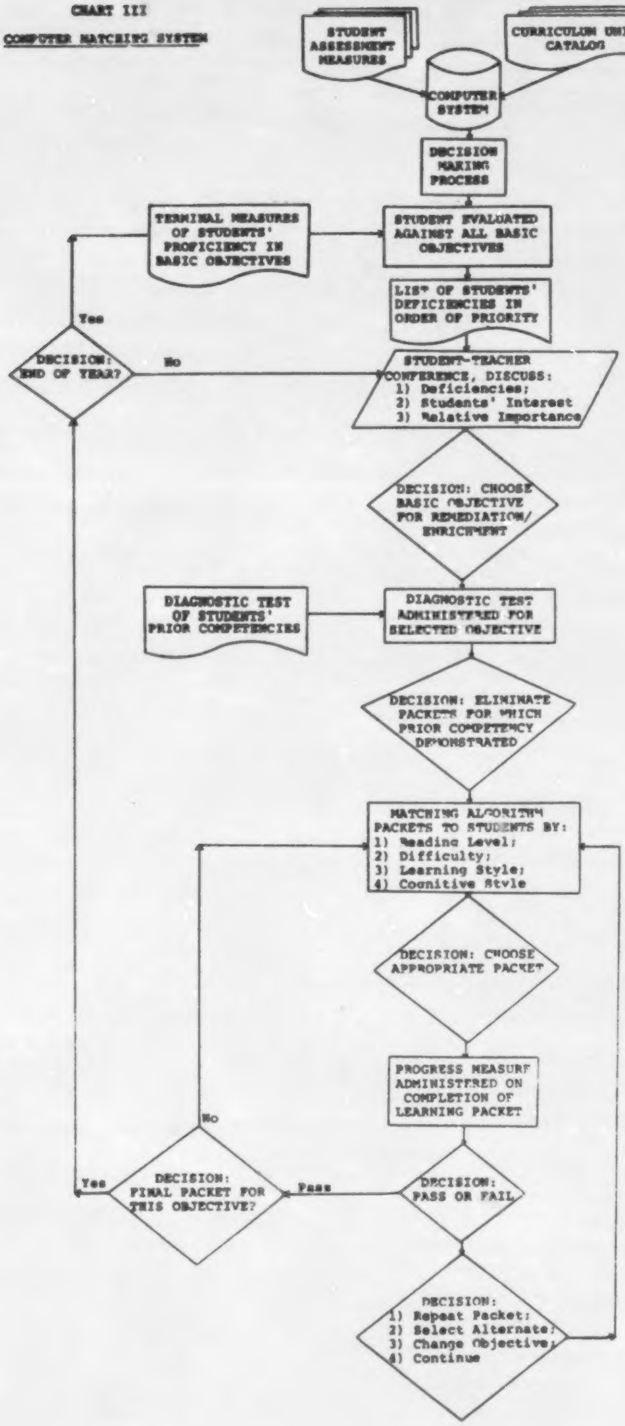
Phase I--A student's record is obtained from the optical scanning and machine scoring of termial measures and aptitude tests. A student record is created from the test information and is stored in a direct-access file. The coded data on curriculum units is used to create a curriculum bank, which is also stored on a disk. The student's deficiencies are arranged by the system according to a predetermined hierarchy of importance. A list of the top 10 deficiencies in a priority order is prepared for each student to permit selection of two objectives for initial concentration.

Phase II--Student reacts directly with the computer and enters the diagnostic test results for the selected objective and indicates that he would like the first assignment. The computer adds the diagnostic data to the student record and accesses the curriculum file for the chosen objective. The actual assignment is then made by means of a complex matching algorithm which selects a learning unit representing the best fit between the student's learning aptitude and the available curriculum treatment.

When the student completes his current assignment he takes a brief progress test and returns to the terminal to indicate success or failure. In the fail case, the computer implements a decision made previously by the

CHART III

COMPUTER MATCHING SYSTEM


STUDENT ASSESSMENT MEASURES
CURRICULUM UNITS CATALOG
COMPUTER SYSTEM
DECISION MAKING PROCESS
TERMINAL MEASURES OF STUDENTS' PROFICIENCY IN BASIC OBJECTIVES
STUDENT EVALUATED AGAINST ALL BASIC OBJECTIVES
LIST OF STUDENTS' DEFICIENCIES IN ORDER OF PRIORITY
Yes
DECISION: END OF YEAR?
No
STUDENT-TEACHER CONFERENCE, DISCUSS: 1) Deficiencies; 2) Students' Interest 3) Relative Importance
DECISION: CHOOSE BASIC OBJECTIVE FOR REMEDIATION/ ENRICHMENT
DIAGNOSTIC TEST OF STUDENTS' PRIOR COMPETENCIES
DIAGNOSTIC TEST ADMINISTERED FOR SELECTED OBJECTIVE
DECISION: ELIMINATE PACKETS FOR WHICH PRIOR COMPETENCY DEMONSTRATED
MATCHING ALGORITHM PACKETS TO STUDENTS BY: 1) Reading Level; 2) Difficulty; 3) Learning Style; 4) Cognitive Style
DECISION: CHOOSE APPROPRIATE PACKET
PROGRESS MEASURE ADMINISTERED ON COMPLETION OF LEARNING PACKET
No
Yes
DECISION: FINAL PACKET FOR THIS OBJECTIVE?
Pass
DECISION: PASS OR FAIL
DECISION: 1) Repeat Packet; 2) Select Alternate; 3) Change Objective; 4) Continue

instructional staff concerning the appropriate remedial action. The system could for example, assign another unit on the same instructional topic with a different learning style, move to the next topic in the instructional sequence, or begin assignments for another objective. If the student has successfully passed the progress test, the computer will search the next learning topic in the instructional sequence for the best assignment. This cycle is repeated throughout the student's learning path with the system tracking the student's progress, branching to remedial paths when needed, and adapting instructional assignments to the student's learning abilities. An interactive language leads the student through the process so that he is not required to memorize any transaction formats or complex data codes.

Phase III--The system takes information from the terminal on a real-time basis and stores it in a direct-access device for overnight processing. Every morning, before the opening of school, the data stored from the previous day is sorted and sequentially run against a series of programs in the background partition. This phase is involved with the preparation of all the reports, such as, summary reports, student reports, and curriculum reports.

DECISION MAKING PROCESS

Chart III shows the flow of the major steps. The first major decision is the choice of the particular objective based on student deficiencies. The next decision point is reached after the student has taken the diagnostic test for the selected objective and the scores have been entered into his evaluation file. The third decision is the selection of the appropriate packet-search for a match between the student's measured learning characteristics and the coded pocket characteristics. The packet choice is based on a matching algorithm consisting of first searching for perfect match to finally assigning what is available, disregarding cognitive style, learning style, reading level, or aptitude level. The next decision point is reached after the student completes the packet assignment, takes the progress test, enters a pass-fail, and receives another packet assignment. When the student has completed the final segment for a particular objective, he is instructed to take the terminal measure. The entire process is then repeated for another objective.

In the five schools where the program is in operation, students are usually scheduled to a center for two fifty-minute periods per week. The center is self-contained, and its activities are managed by a trained teacher and an aide. A number of problems have been faced relating to the time needed for testing, teacher training, defining the objectives, selecting and developing measures for these objectives, and carefully structuring all the components in order to measure and remediate on a very specific level. Changes have been made since 1971, when the system was first put in use in one school, based on comments which have been raised on the reliability of the instruments and the techniques used for assessing student needs and on results of student performance. This system offers a strategy for adapting instruction to match teaching objectives with the individual learning needs of each student.

CONCLUSIONS

A basic assumption is made that individualized, mastery-oriented instruction requires the capability of computers to test students, help keep track of their deficiencies and accomplishments, and provides for each student the material appropriate to his/her needs.

It has been demonstrated that administration of tests by the computer requires 10-20% less time than does hand scoring. Instructors then use the time, normally spent scoring tests, providing students with test results, and working with individual students. There are problems, however, which have been encountered. Off-line tests have been often circulated among the students. To prevent this occurrence, many equivalent and nonidentical tests must be prepared. Questions have arisen as to the number of test questions needed to determine mastery for each behavioral objective. Since students must be tested frequently, lack of sufficient equipment has created difficult situations. Inability to use graphics, especially moving graphics, also has hindered the kinds of questions which can be asked. Restrictions that have caused the almost exclusive use of true or false and multiple choice questions have been difficult to deal with. Even where free form questions are permitted, there are constraints which disallow freedom of expression.

"Tailored Testing" has been discussed by theoreticians and researchers as an effective application of latent trait theory. It is believed it can be best implemented in a computer-interactive method where, after a test item appears, and the answer is given, the next item will be more difficult, if the answer given is correct. If the answer is not correct, an easier item will follow. With each response, the computer makes a revised estimate of the examinee's ability. Each succeeding estimate then becomes more reliable. The test is terminated when the estimate reaches a specified level of reliability.

For example: If nothing is known about a particular student, the best prior estimate of ability is the standardized mean of the population of students (0.00). Before testing begins, the standard error of this estimate is the standard deviation of standard scores, or 1.000. After each item is presented, attempted, scored, the standard error of the estimate of ability decreases. Items are selected for presentation in the order that will decrease this standard error more rapidly. It has been found that there is a dramatic reduction in the number of items required for valid measurement of those making tailored testing economically feasible. But this does depend on excellent ability banks. the U.S. Civil Service Commission is moving toward the implementation of a nationwide computer assisted testing system by 1980, for Federal positions. It is expected that job applicants will be able to be tested wherever a terminal can be installed; shortened testing time, since fewer items will be necessary, and immediate reports on test scores, weighted for relevance to various occupations and related to current availability of jobs.

Educators could certainly benefit from tailored testing. Computerized guidance and counseling, which is presently in operation, would be greatly improved as realistic educational goals and career choices could be attained. Also, since the measurement problem is the heart of computer-managed instruction, and consequently for true individualization of instruction, tailored testing is bound to have important implications for computer managed instruction.

PROJECT TORQUE

Judah L. Schwartz & Edwin F. Taylor
Division for Study and Research in Education
Massachusetts Institute of Technology
and
Project TORQUE, Education Development Center

Project TORQUE* is a research and development effort concerned with alternatives to the present assessment instruments in the elementary schools. The first task has been mathematics and in particular, the skills of computation and measurement. The computation package is complete and has been field tested with 15,000 students during the 1978-79 school year. The measurement package was completed in the summer of 1979. Each package contains tests, a diagnostic manual for teachers, an administrative manual for teachers and administrators, and a management package of record-keeping and report forms. The measurement package also includes games and activities for classroom use and for closer diagnosis of student difficulties.

We believe the following to be true of TORQUE materials:

1. They intertwine testing and instruction. Criterion-referencing, immediate feedback, and diagnostic usefulness based on constructed answers provide teachers with tools of instruction and all interested parties with a common language about the subject matter.

2. They are based on a detailed analysis of children's behavior in the skills being tested. The tests are developed and validated by observing children perform on specially designed games and in "real life" activities, by watching children do the tests, and by analyzing large numbers of answers.

3. They can satisfy a spectrum of purposes, from formal testing through fine-grained diagnosis of difficulties to steps for overcoming the difficulties and elements of curriculum. Because the materials can be separated by the skill being assessed, they can be useful in demonstrating that students have met applicable school district or state competency requirements.

4. The procedures by which TORQUE tests are developed and validated are applicable to subject matter other than mathematics and school levels other then elementary schools.

COMPUTATION

There are two kinds of computation assessment materials:

1. Self-checking sheets in which computation begins with an arbitrary number supplied either by the child or by the teacher and which leads back to the same number after computation is carried out. (See Figures 1a, 1b)

---

*Project TORQUE is supported by grants from the Carnegie Corporation of New York and the Ford Foundation.

2. More formal computation tests on ordering, addition, subtraction, multiplication, and division of integers, fractions, and decimals. This second kind of test consists of 30 one-page No-Carbon-Required (NCR) forms that provide immediate feedback to children and diagnostic help to teachers. (See Figures 2a, 2b)

TORQUE COMPUTATION TESTS ARE:

- Constructed-answer: Students solve the exercises and write down the answers themselves, demonstrateing their ability to do the problem rather than their ability to select a correct answer. The constructed answers are very much more helpful for diagnosing student difficulties than are multiple-choice answers.

- Diagnostically useful: The second sheet on the NCR form gives the teacher the student's answers, the correct answers, a general description of the kinds of errors most often found on the test, and specific wrong answers that may indicate these errors. Supplementing the information on each test is a Diagnostic Manual that helps the teacher understand more comprehensively the kinds of errors students make. The error categories are derived from the analysis of thousands of student responses.

- Immediate-feedback: After the two sheets of these NCR forms are separated by the teacher, the first sheet is returned to the student who, by folding over the right-hand edge, can look at the correct answers immediately. The back of the student sheet contains a statement of the purpose of that particular test. The student can take the test home to inform parents. The teacher can use the diagnosis on the second sheet to help the student right away. The self-checking sheets yield the same number at the end that the student or teacher inserted at the beginning, so the correctness can be judge immediately.

- Criterion-referenced: Each computation test is interpreted in terms of the structure of the computation task itself. Performance is judged according to competence, rather than according to comparison with other children.

- Ramped: Each computation test deals with a single topic (e.g., multiplication of integers) and the items on each test progress in difficulty so that the student and teacher can determine the point at which the student needs help.

- Flexible to use: Self-checking sheets can be used as classroom or individual activities, with the child providing the initial number; later they can serve as tests, with the teacher supplying the initial number. Computation tests can be given to groups of students, entire classrooms, or individual students. Appropriate tests can be administered once or twice a year in alternative ways tailored to the needs of the students, the teacher, and the school.

- Helpful in reporting to various audiences: Students know immediately that they have demonstrated skills; parents receive the student's work and a statement of the purpose of each test; teachers can identify student difficulties and deal with them individually or in small groups; a management package for record-keeping and reporting assists the teacher in organizing the results and provides summary data to the school administration.

TORQUE COMPUTATION TESTS ARE:

- NOT secret: Self-checking sheets require computation with an initial number provided by the student or the teacher. Only when the teacher supplies the number does it become a test; therefore the sheets themselves may be made public without reducing the effectiveness of their testing function. Five alternative equivalent versions of each of the 30 computation tests are available (and more may easily be generated) thus eliminating the need for secrecy for these tests. Because secrecy is not required for TORQUE computation tests, every member of the community with an interest in the schools can see in full detail the assessment package by which school performance is being measured. Legal requirements for disclosure to parents are automatically met by sending home student NCR sheets and completed self-checking sheets.

- NOT norm-referenced: TORQUE tests do not compare one student with other students either locally or nationally. Each school system can decide which computation skills are important and what performance will be considered adequate for key grades.

- NOT timed: We feel that tests should assess how well students perform, not how quickly they perform or how they react to the pressure of time. In contrast, norm-referenced tests are given under "standard" circumstances, including the time allowed for each section of the test.

One long-range benefit to those who use TORQUE materials is a sharpened perception of all computation. We have found that teachers, students, parents, and administrators can share a language about computation problems and computation errors whose usefulness may equal that of the tests themselves.

MEASUREMENT

Project TORQUE has developed tests and associated activities that assess competence in the measurement of length, time, area, weight, and volume. These skills are an important part of the application of mathematics to quantitative aspects of everyday life. The major move toward metric measurement in many school systems is strongly reinforcing the teaching of measurement skills. All TORQUE measurement tests use metric units.

TORQUE measurement tests are developed within the framework of an empirically verified task analysis of the measurement of extensive magnitudes. This analyis accounts for the common structure of the measurement of length, area, volume, weight and time.

TORQUE measurement tests are validated according to a new procedure in which the performance of children on the tests are verified to be similar to their performance on games and other "real-life" activities that allow detailed observation of performance.[1] For classroom use, the games and other activities provide not only experience in measurement for students but also a more personal and detailed observation mechanism for the teacher (See Figure 3).

A total of 7 measurement tests have been developed for the five skills (See Figure 4). The written tests are printed on No-Carbon-Required (NCR) forms so that students and teachers can both keep a copy, the student can correct the test immediately, and the most common errors can be provided to the teacher for diagnostic use. A Diagnostic Manual, an Administrative Manual, and record-keeping forms for individual students and whole classes are also available.

To complete the measurement package, we have developed a summary manipulative test that combines several measurements. A die-cut cardboard cut-out can be folded into a rectangular parallelpiped on which measurements can be made of lengths, areas, and volume. Partial and total filling with sand yields several weights. Such a test is more than a group of separate measurement tests in a more "real" setting but also requires students to select the correct attribute (length, area, etc.) for each measurement.

Like the TORQUE computation materials, the measurement materials do not need to be secret. Each test is available in multiple alternative but equivalent forms, so that the student can keep a copy of the test and take it home for examination by parents. The games and other associated activities can be entirely public and can be marketed outside of schools.

FUTURE WORK

With the completion of the computation and measurement materials TORQUE has turned to the development of materials in estimation and word problems.[2]

Pilot work is currently underway in loose collaboration with colleagues in psychology, linguistics and mathematics at the University of California in San Diego, the University of Pittsburgh, the University of Puerto Rico and the Massachusetts Institute of Technology. As was the case with its earlier effort TORQUE is continuing to involve teachers and children in all stages of the research and development process.

References

[1] Schwartz, J.L., and Taylor, E.F. Valid Assessment of Complex Behavior; The TORQUE approach in Quarterly Newsletter of the Institute for Comparitive Human Development, The Rockefeller University, N.Y., Vol. 2, No. 3, July 1978.

[2] See for example:

a. Lebesque, H., Sur La Mesure des Grandeurs, Enseignment Mathematique, 31-34, 1933-1936. Reprinted in translation in Measure and The Integral, K.O. May (ed.), San Francisco, Holden Day 1966.

b. Schwartz, J.L., Semantic Aspects of Quantity, MIT Division for Study & Research in Education, Unpublished monograph, May 1976.

c. Whitney, H. The Mathematics of Physical Quantities Part I; Mathematical Models for Measurement, American Mathematical Monthly, Vol. 75, No. 2, 1968.

Whitney, H. The Mathematics of Physical Quantities Part II; Quantity Structures and Dimensional Analysis, American Mathematical Monthly, Vol 75, No. 3, 1968.

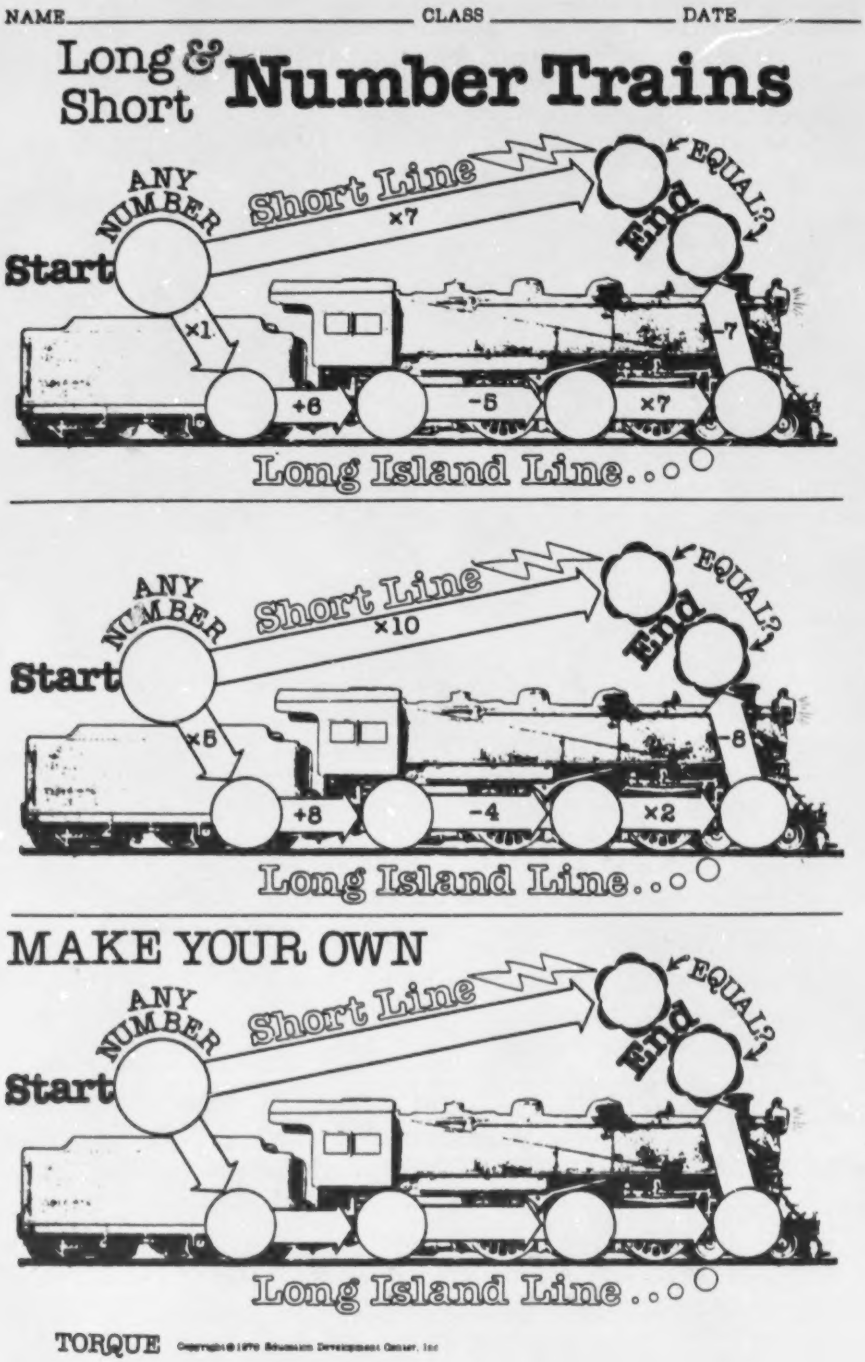
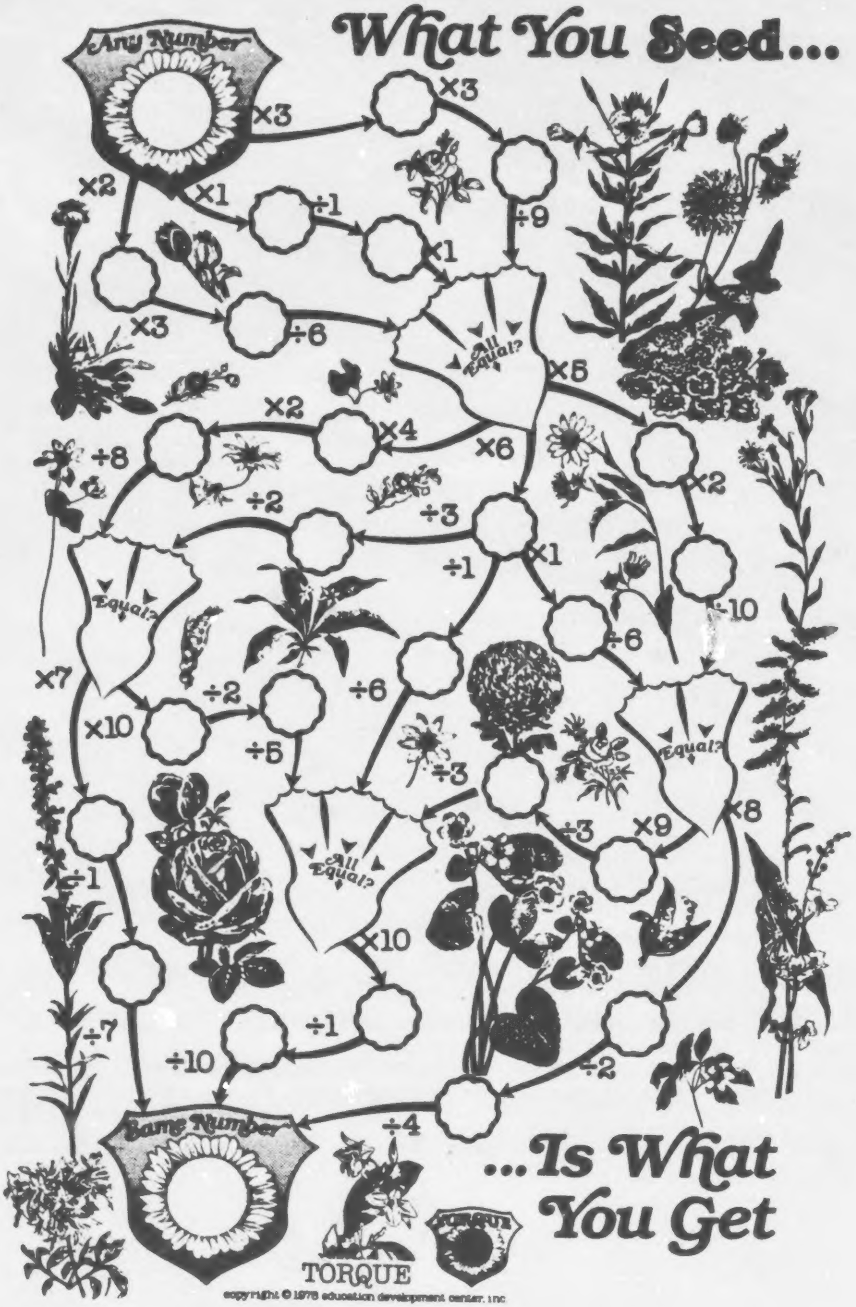
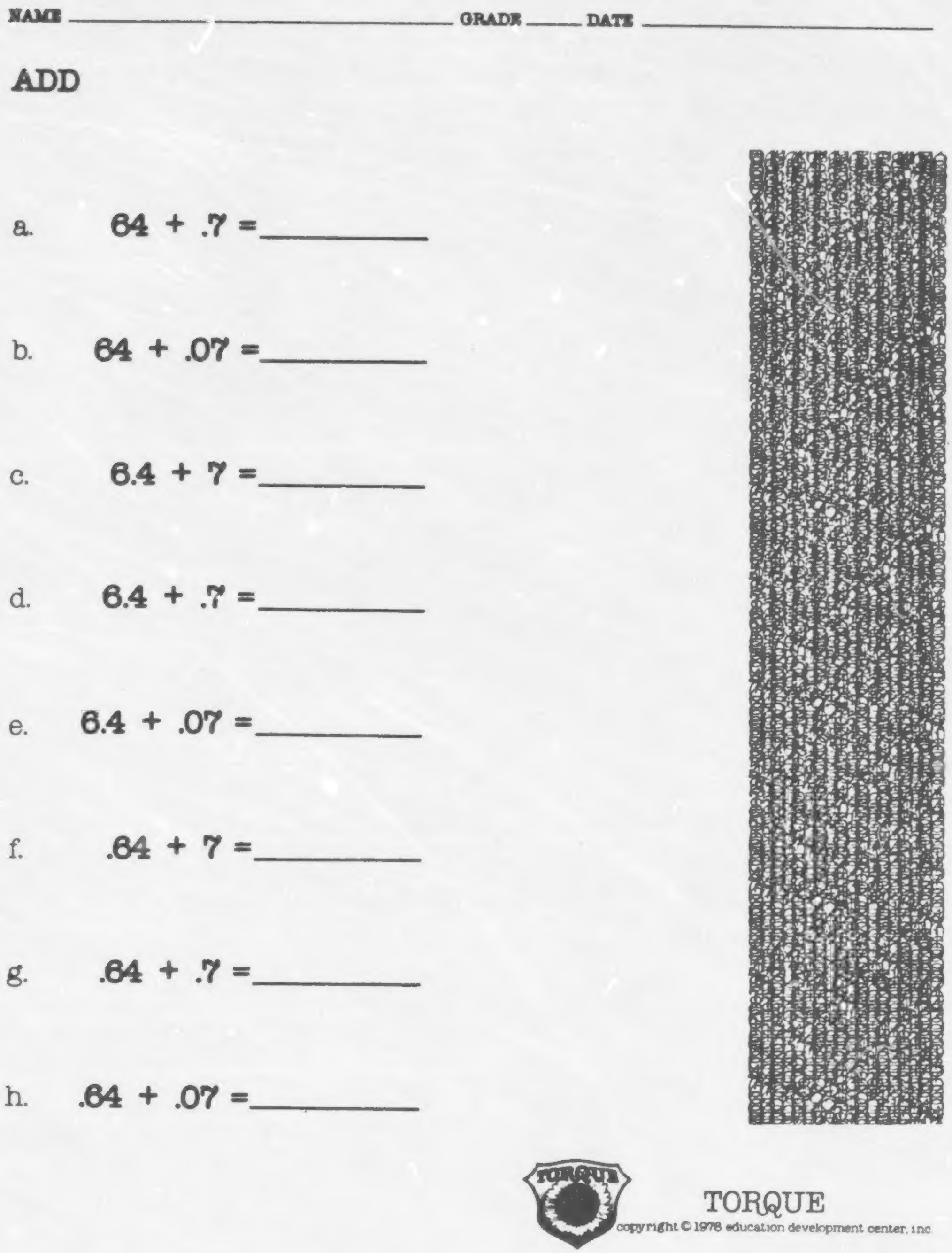

NAME
CLASS
DATE
Long & Short Number Trains
ANY NUMBER
Start
Short Line
×7
EQUAL?
End
×1
+6
-5
×7
-7
Long Island Line
ANY NUMBER
Start
Short Line
×10
EQUAL?
End
×5
+8
-4
×2
-8
Long Island Line
MAKE YOUR OWN
ANY NUMBER
Start
Short Line
EQUAL?
End
Long Island Line
TORQUE

Fig. 1a Self-checking sheet for use with lower grades (no division required).


What You Seed...
Any Number
×3
×3
×2
×1
÷1
×1
÷9
×3
÷6
All Equal?
×5
×2
×4
×6
÷8
÷2
÷3
×2
×1
÷1
Equal?
÷10
÷6
×7
÷2
÷6
×10
÷5
÷3
Equal?
×8
÷3
×9
All Equal?
×10
÷1
÷1
÷7
÷10
÷2
÷4
Same Number
...Is What You Get
TORQUE

copyright © 1976 education development center, inc

Fig. 1b Self-checking (with division) that is particularly useful with fractions and decimals.



NAME ______________________________ **GRADE** _____ **DATE** ____________________

ADD

a. 64 + .7 = __________

b. 64 + .07 = __________

c. 6.4 + 7 = __________

d. 6.4 + .7 = __________

e. 6.4 + .07 = __________

f. .64 + 7 = __________

g. .64 + .7 = __________

h. .64 + .07 = __________

[redacted]

TORQUE

copyright © 1978 education development center, inc

Fig. 2a Student's Copy of Decimal Addition Test

NAME ______________________ GRADE _____ DATE ______________

SUBTRACT

a.	763 -453	b.	820 -600
c.	525 -240	d.	524 -306
e.	659 -287	f.	674 -197
g.	448 -359	h.	381 -96
i.	900 -540	j.	407 -260
k.	703 -476	l.	500 -138

Most common errors arise from:

- adding instead of subtracting (add)
- subtracting the smaller digit from the larger regardless of position (switch digits)
- forgetting that a column has been borrowed from (forget borrow)

a. 1216—add
b. 1420—add

c. 765—add
325—switch digits
385—forget borrow
d. 830—add
222—switch digits
228-forget borrow

e. 946—add
43[illegible]—switch digits
[illegible]72—forget borrow
f. 871—add
523—switch digits
487, 587, 577 } forget borrow

g. 807—add
111—switch digits
99, 189, 199 } forget borrow
h. 477—add
315—switch digits
295, 385, 395 } forget borrow

i. 1440—add
440—switch digits
460—forget borrow

j. 667—add
267—switch digits
247—forget borrow

k. 1179—add
267—switch digits
237, 327, 337 } forget borrow

l. 6[illegible]—add
4[illegible]—switch digits
372, 462, 472 } forget borrow

ANSWERS

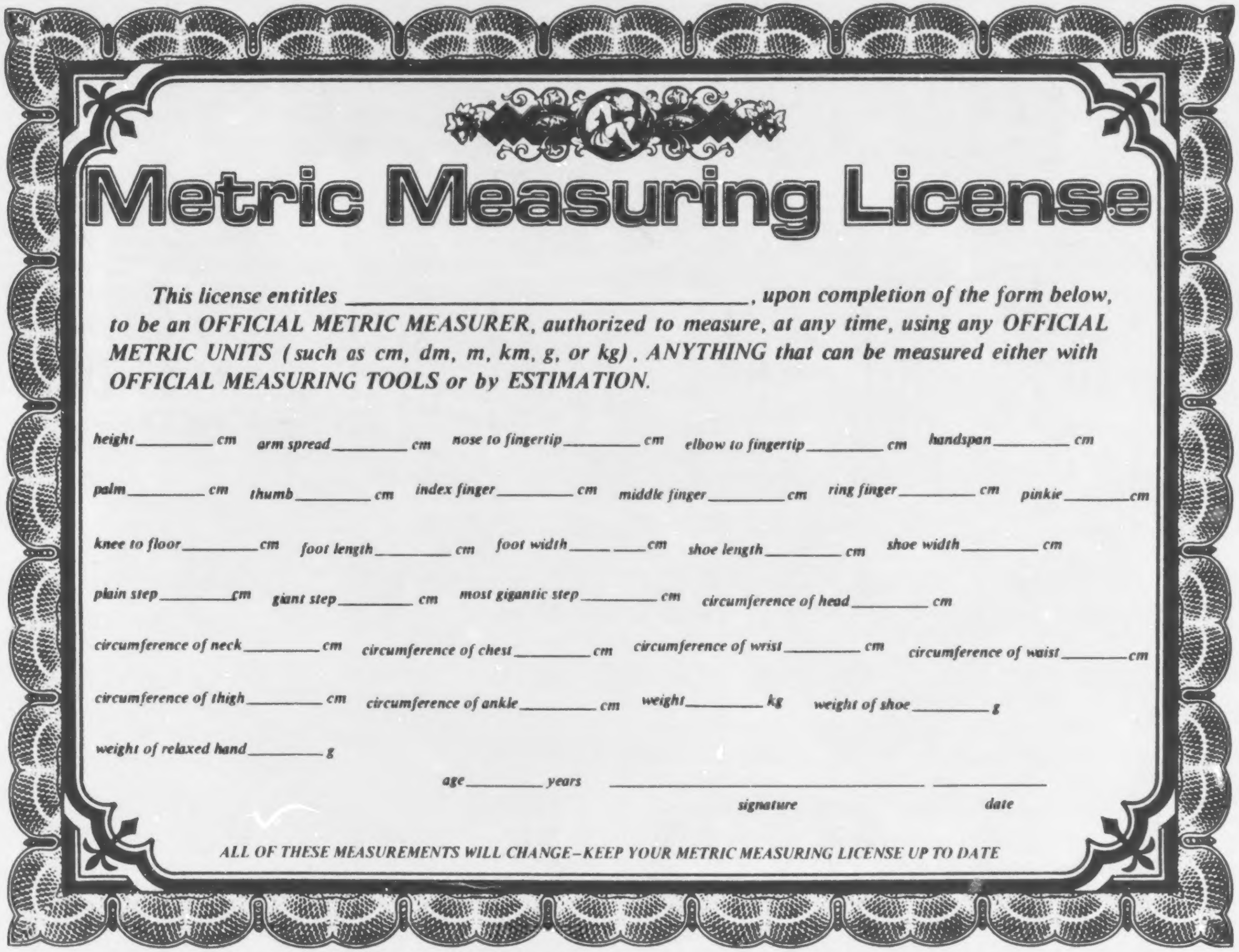
a. 310
b. 220
c. 285
d. 218
e. 372
f. 477
g. 89
h. 285
i. 360
j. 147
k. 227
l. 362

TORQUE

copyright © 1978 education development center, inc.

IST

Fig. 2b Teacher's Copy of Integer Subtraction Test



Metric Measuring License

This license entitles ______________________________, upon completion of the form below, to be an OFFICIAL METRIC MEASURER, authorized to measure, at any time, using any OFFICIAL METRIC UNITS (such as cm, dm, m, km, g, or kg), ANYTHING that can be measured either with OFFICIAL MEASURING TOOLS or by ESTIMATION.

height _______ cm arm spread _______ cm nose to fingertip _______ cm elbow to fingertip _______ cm handspan _______ cm

palm _______ cm thumb _______ cm index finger _______ cm middle finger _______ cm ring finger _______ cm pinkie _______ cm

knee to floor _______ cm foot length _______ cm foot width _______ cm shoe length _______ cm shoe width _______ cm

plain step _______ cm giant step _______ cm most gigantic step _______ cm circumference of head _______ cm

circumference of neck _______ cm circumference of chest _______ cm circumference of wrist _______ cm circumference of waist _______ cm

circumference of thigh _______ cm circumference of ankle _______ cm weight _______ kg weight of shoe _______ g

weight of relaxed hand _______ g

age _______ years ____________________ __________

signature date

ALL OF THESE MEASUREMENTS WILL CHANGE–KEEP YOUR METRIC MEASURING LICENSE UP TO DATE

Copyright © 1974 by Education Development Center, Inc.

Fig. 3 An example of TORQUE classroom materials in measurement.

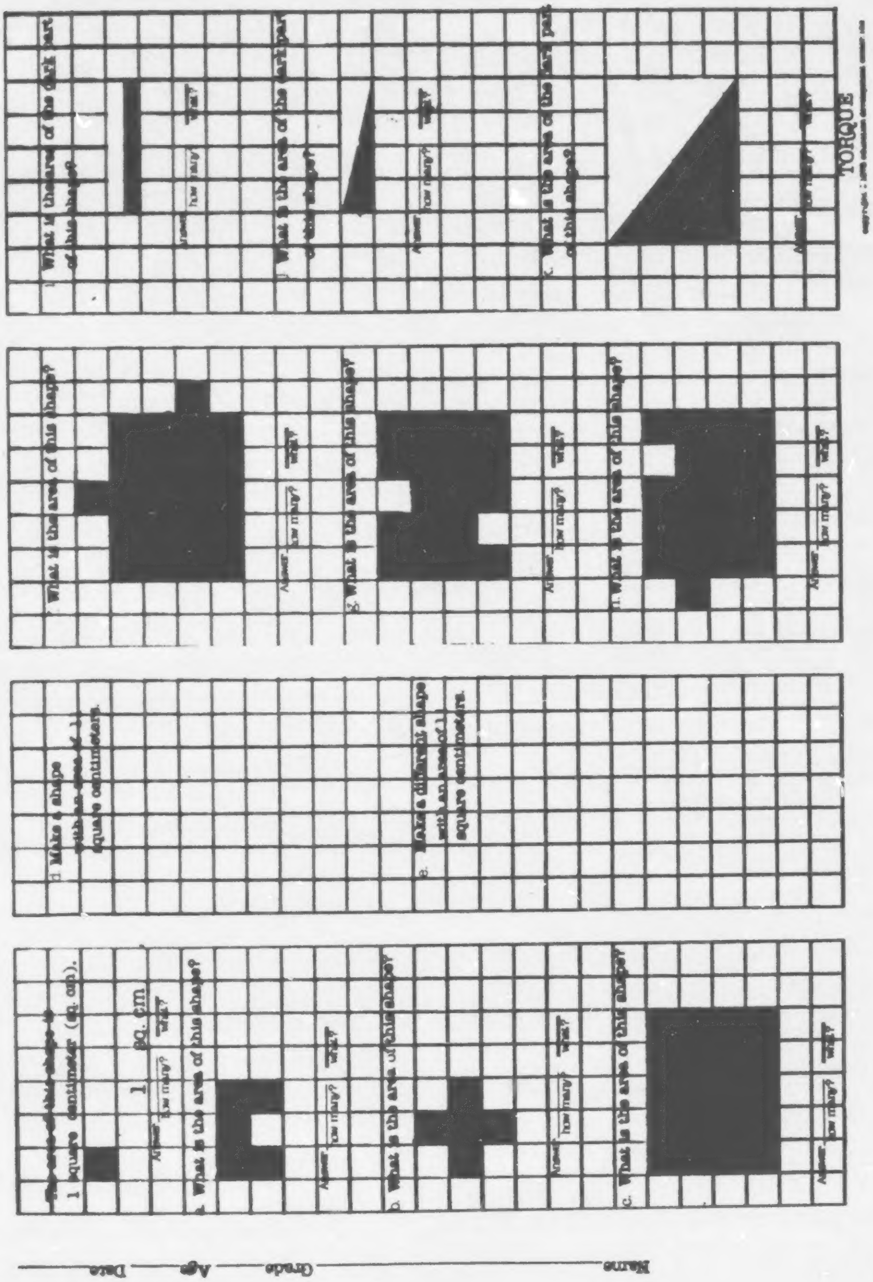

The area of this shape is
1 square centimeter (sq. cm).
Answer 1 sq. cm
how many? what?
a. What is the area of this shape?
b. What is the area of this shape?
c. What is the area of this shape?
d. Make a shape with an area of 1 square centimeters.
e. Make a different shape with an area of 1 square centimeters.
What is the area of this shape?
g. What is the area of this shape?
h. What is the area of this shape?
What is the area of the dark part of this shape?
What is the area of the dark part of this shape?
k. What is the area of the dark part of this shape?
Answer how many? what?
TORQUE
Name Grade Age Date

Fig. 4 Preliminary version of a test of area measurement.

COGNITIVE SCIENCE, 2, 155-192 (1978)

Diagnostic Models for Procedural Bugs in Basic Mathematical Skills*

JOHN SEELY BROWN AND RICHARD R. BURTON

Bolt Beranek and Newman

A new diagnostic modeling system for automatically synthesizing a deep-structure model of a student's misconceptions or bugs in his basic mathematical skills provides a mechanism for explaining *why* a student is making a mistake as opposed to simply identifying the mistake. This report is divided into four sections: The first provides examples of the problems that must be handled by a diagnostic model. It then introduces *procedural networks* as a general framework for representing the knowledge underlying a skill. The challenge in designing this representation is to find one that facilitates the discovery of misconceptions or bugs existing in a particular student's encoding of this knowledge. The second section discusses some of the pedagogical issues that have emerged from the use of diagnostic models within an instructional system. This discussion is framed in the context of a computer-based tutoring/gaming system developed to teach students and student teachers how to diagnose bugs strategically as well as how to provide a better understanding of the underlying structure of arithmetic skills. The third section describes our uses of an executable network as a tool for *automatically* diagnosing student behavior, for automatically generating "diagnostic" tests, and for judging the diagnostic quality of a given exam. Included in this section is a discussion of the success of this system in diagnosing 1300 school students from a data base of 20,000 test items. The last section discusses future research directions.

> If you can both listen to children and accept their answers not as things to just be judged right or wrong but as pieces of information which may reveal what the child is thinking you will have taken a giant step toward becoming a master teacher rather than merely a disseminator of information.
>
> J. A. EASLEY, JR. & R. E. ZWOYER (1975)

INTRODUCTION

One of the greatest talents of teachers is their ability to synthesize an accurate "picture," or model, of a student's misconceptions from the meager evidence

---

*This paper is a substantially expanded version of a paper written with Kathy M. Larkin entitled "Representing and Using Procedural Bugs for Educational Purposes," which appears in *Proceedings of the ACM*, National Conference, ACM 77, October 1977.

Development of the general framework of Diagnostic Models which underlies this research was supported, in part, by the Advanced Research Projects Agency, Air Force Human Resources Laboratory, Army Research Institute for Behavioral and Social Sciences, and Navy Personnel Research and Development Center under Contract No. MDA903-76-C-0108.

Requests for reprints should be addressed to Richard R. Burton, Bolt Beranek and Newman Inc., 50 Moulton Street, Cambridge, Mass. 02138.

inherent in his errors. A detailed model of a student's knowledge, including his misconceptions, is a prerequisite to successful remediation. The structure, use, and *inference* of such models for procedural skills in mathematics is the topic of this paper. In particular we shall describe some initial efforts in the development and use of a representational technique called "procedural networks" as the framework for constructing *diagnostic models*—i.e., models that capture a student's common misconceptions or faulty behavior as simple changes to (or mistakes in) a correct model of the underlying knowledge base. By being able to synthesize such deep-structure diagnostic models *automatically*, we can provide both a teacher and an instructional system with not only an identification of *what* mistakes a student is making, but also an explanation of *why* those mistakes are being made. Such a system also has profound implications for testing, since a student need no longer be evaluated solely on the number of errors appearing on his test, but rather on the fundamental misconceptions which he harbors.

This paper consists of four sections. The first provides examples of the problems that must be handled by a diagnostic model. It then introduces procedural networks as a general framework for representing the knowledge underlying a skill. The challenge here is to design a representation that facilitates the discovery of misconceptions or bugs existing in a particular student's encoding of this knowledge. The second section discusses some of the pedagogical issues that have emerged from our use of diagnostic models within an instructional system. This system is a computer-based tutoring game developed to teach both students and student teachers about the strategic diagnosis of bugs. The third section describes our uses of procedural network as a tool for automatically diagnosing student behavior, for automatically generating "diagnostic" tests, and for judging the diagnostic quality of a given exam. Included in this section is a discussion of the success of this system in diagnosing 1300 grade-school students from a data base of 20,000 test items. The last section discusses some future research directions.

1. DIAGNOSTIC MODELS OF BASIC SKILLS

The issues addressed in this paper arose from an investigation of the procedural skills necessary to solve high school algebra problems. These skills include not only the generally recognized rules of algebra, but also such normally implicit skills as the reading of formulas, the parsing of expressions, and the determination of which rules to apply next (Brown & Burton, 1975; Brown, Collins, & Harris, 1978; Matz, 1978). For this paper, however, we limit our discussion to examples encompassing arithmetic skills so that we can concentrate on the critical ideas of diagnosis without the need for a large number of algebraic rules. Limiting our examples to arithmetic also provides a compelling demonstration of the difficulty of the diagnostic task; it clearly demonstrates how much more difficult it is to diagnose what is wrong with a student's method of performing a task—to form a diagnostic model—than it is to perform the task itself. In particular, it is no great challenge to add or subtract two numbers, but, as we shall see, diagnosing misconceptions in these same skills can be quite subtle.

Let us consider diagnosing what is wrong with the arithmetic skills (procedures) of a couple of students. We shall start with a case study in which we examine five "snapshots" of a student's performance doing addition as might be seen on a homework assignment. Before proceeding, look at the following snapshots and try to discover the student's bug.

Sample of the student's work:

$$\begin{array}{r} 41 \\ +9 \\ \hline 50 \end{array} \qquad \begin{array}{r} 328 \\ +917 \\ \hline 1345 \end{array} \qquad \begin{array}{r} 989 \\ +52 \\ \hline 1141 \end{array} \qquad \begin{array}{r} 66 \\ +887 \\ \hline 1053 \end{array} \qquad \begin{array}{r} 216 \\ +13 \\ \hline 229 \end{array}$$

Once you have discovered the bug, try testing your hypothesis by "simulating" the buggy student so as to predict his results on the following two test problems.

$$\begin{array}{r} 446 \\ +815 \\ \hline \end{array} \qquad \begin{array}{r} 201 \\ +399 \\ \hline \end{array}$$

The bug is really quite simple. In computer terms, the student, after determining the carry, forgets to reset the "carry register" to zero and hence the amount carried is accumulated across the columns. For example, in the student's second problem (328 + 917 equals 1345) he proceeds as follows: 8+7=15 so he writes 5 and carries 1; 2+1=3 plus the 1 carry is 4; lastly, 3+9=12 but that 1 carry from the first column is still there—it has not been reset—so adding it in to this column gives 13. If this is the bug, then the answers to the test problems will be 1361 and 700. This bug is not so absurd when one considers that a child might use his fingers to remember the carry and forget to bend back his fingers, or counters, after each carry is added.

A common assumption among teachers is that students do not follow procedures very well and that erratic behavior is the primary cause of a student's inability to perform each step correctly. Our experience has been that students are remarkably competent procedure followers, but that they often follow *the wrong procedures*. One case encountered last year is of special interest. The student proceeded through a good portion of the school year with his teacher thinking that he was exhibiting random behavior in his arithmetic performance. As far as the teacher was concerned there was no systematic explanation for his errors. Here is a sample of his work:

$$\begin{array}{r} 7 \\ +8 \\ \hline 15 \end{array} \qquad \begin{array}{r} 9 \\ +5 \\ \hline 14 \end{array} \qquad \begin{array}{r} 8 \\ +3 \\ \hline 11 \end{array} \qquad \begin{array}{r} 6 \\ +7 \\ \hline 13 \end{array} \qquad \begin{array}{r} 8 \\ +8 \\ \hline 16 \end{array} \qquad \begin{array}{r} 9 \\ +9 \\ \hline 18 \end{array} \qquad \begin{array}{r} 17 \\ +8 \\ \hline 25 \end{array} \qquad \begin{array}{r} 19 \\ +4 \\ \hline 23 \end{array} \qquad \begin{array}{r} 87 \\ +93 \\ \hline 11 \end{array}$$

$$\begin{array}{rrrrr} 365 & 679 & 923 & 27{,}493 & 797 \\ +574 & +794 & +481 & +1{,}509 & +48{,}632 \\ \hline 819 & 111 & 114 & 28{,}991 & 48{,}119 \end{array}$$

There is a clue to the nature of his bug in the number of ones in his answers. Every time the addition of a column involves a carry, a 1 mysteriously appears in that column; he is simply writing down the carry digit and forgetting about the units digit! One might be misled by 17+8 which normally involves a carry yet is added correctly. It would seem that he is able to do simple additions by a completely different procedure—possibly by counting up from the larger number on his fingers.

The manifestation of this student's simple bug carries over to other types of problems which involve addition as a subskill. What answer would he give for the following?

> A family has traveled 2975 miles on a tour of the U.S. They have 1828 miles to go. How many miles will they have traveled at the end of their tour?

He correctly solved the word problem to obtain the addition problem 2975 + 1875, to which he answered 3191. Since his work was done on a scratch sheet, the teacher only saw the answer which is, of course, wrong. As a result, the teacher assumed that he had trouble with word problems as well as with arithmetic.

When we studied this same student's work in other arithmetic procedures, we discovered a recurrence of the same bug. Here is a sample of his work in multiplication:

$$\begin{array}{rrrrr} 68 & 734 & 543 & 758 & 2764 \\ \times 46 & \times 37 & \times 206 & \times 296 & \times 53 \\ \hline 24 & 792 & 141 & 144 & 2731 \end{array}$$

Several bugs are manifested here, the most severe one being that his multiplication algorithm mimics the column behavior of his addition algorithm. But notice that the bug in his addition algorithm is also present in his multiplication procedure. The "determine-unit-and-carry" subprocedure bug shows up in both his multiplication and addition. For example, to do 68×46, in the first column he performs 8×6, gets 48 and then writes down the "carry," in this case 4, ignoring the units digit. Then, he multiplies 6×4 to get 2 for the second column. All along he has a complete and consistent procedure for doing arithmetic. His answers throughout all of his arithmetic work are far from random; in fact, they display near perfection with respect to *his* way of doing it.

A First Approximation to Representing Procedural Skills

For a computer system to be capable of diagnosing aberrant behavior such as the above, the procedural skill being taught must be represented in a form amenable to modeling *incorrect* as well as correct subprocedures of that skill. Furthermore, the representation must allow the intermixing of both the correct and incorrect subskills, so that the model can capture those parts of a skill that are correct as well as those that are wrong. The breakdown of the skill into shared subskills can also account for the recurrence of similar errors in different skills. We introduce the term *diagnostic model* to mean a representation of a student's procedural knowledge or skill that depicts his internalization of a skill as a variant of a correct version of that skill.

In addition to satisfactory representational techniques, the diagnostic modeling task requires that the representation of a particular correct skill make explicit much of the tacit knowledge underlying the skill. In particular, the correct model must contain *all* of the knowledge that can possibly be misunderstood by the student or else some student misconceptions will be beyond the diagnostic modeling capabilities of the system. For example, if the model of addition does not include the transcription of the problem, the system would never be able to diagnose a student whose bug is to write 9s that he later misreads as 7s.

The technique we use to represent diagnostic models is a *procedural network*.[1] A procedural network model for a correct skill consists of a collection of procedures with annotations in which the control structure (i.e., calling relationships) between procedures are made explicit by appropriate links. Each procedure node has two main parts: a conceptual part representing the intent of the procedure, and an operational part consisting of methods for carrying out that intent. The methods (also called implementations) are programs that define how the results of other procedures are combined to satisfy the intent of a particular procedure.[2] Any procedure can have more than one implementation, thus providing a way to model *different methods* for performing the same skill. For most skills, the network representation takes the form of a lattice. Figure 1 presents a partial breakdown of a portion of the addition process into a procedural network. Conceptual procedure nodes are enclosed in ellipses. The top procedure in the lattice is addition.[3] Two of the possible algorithms for doing addition are presented as

---

[1] This term has been used by Earl Sacerdoti (1977) to describe an interesting modeling technique for a partially ordered sequence of annotated steps in a problem solving "plan" as well as for specifying control information. Our use of procedural nets coincides with his on this latter feature but differs from and is less developed than his with regard to "plans."

[2] The language we have used to define these programs is LISP. The particular programming language is unimportant from a theoretical standpoint because an implementation is nonintrospectable. The modeling aspects of the network must occur at the conceptual procedure level. For example, the implementation of the subtraction facts table look-up procedure in the computer is necessarily different from that in the student. However, the conceptual properties of the two implementations can be made to agree to an "appropriate" level of detail. Those aspects which are appropriate for our task—the invoking of other procedures, the values returned, the relevant side effects—are included in the network, while the implementation details that may differ are "swept under the rug" into the program. This distinction between conceptual and implementation details also allows skills to be modeled efficiently at different levels which may be appropriate for different tasks.

[3] This simplified representation demonstrates only those features of the procedural network particularly relevant to the diagnostic task. The actual breakdown into subprocedures may be different in a particular network and will be considerably more detailed.

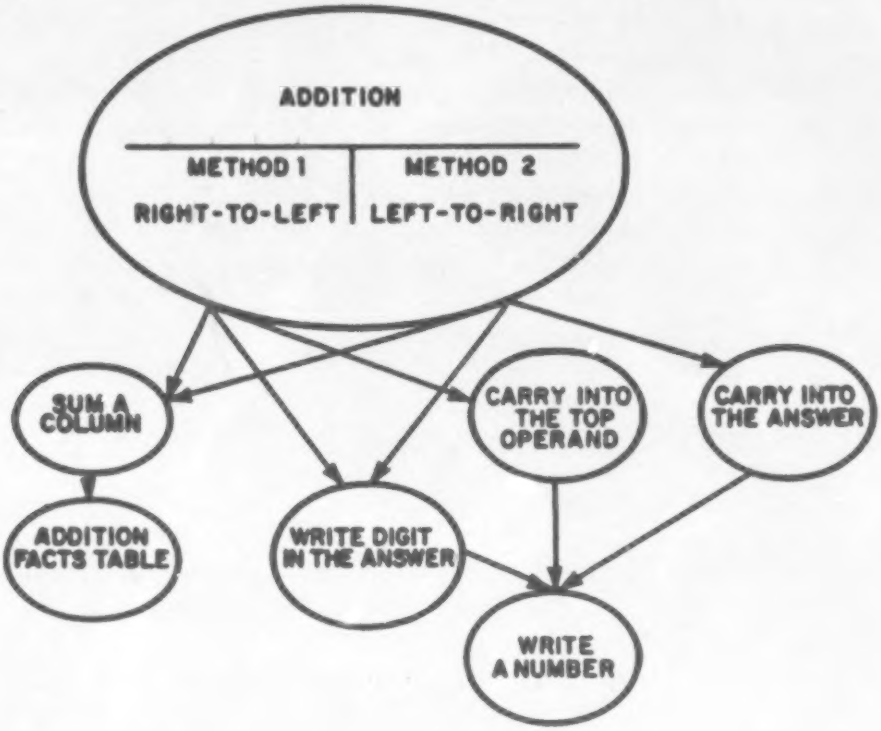

ADDITION
METHOD 1
RIGHT-TO-LEFT
METHOD 2
LEFT-TO-RIGHT
SUM A COLUMN
CARRY INTO THE TOP OPERAND
CARRY INTO THE ANSWER
ADDITION FACTS TABLE
WRITE DIGIT IN THE ANSWER
WRITE A NUMBER

FIG. 1 A portion of the procedural network for addition.

alternative methods under the conceptual node for addition. In Method 1, the standard algorithm, the columns are added from right to left with any carries being written above and included in the column sum of the next column to the left. In Method 2, the columns are added from left to right with any carries being written below the answer in the next column to the left. If any carries occur in the problem, they must be added in a second addition. Notice that these two methods share the common procedures for calculating a column sum and writing a digit in the answer, but differ in the procedure they use when carrying is necessary. One structural aspect of the network is to make explicit any subprocedures that can be potentially shared by several higher level procedures.

The decomposition of a complex skill into all of its conceptual procedures terminates in some set of primitives that reflect assumed elements of an underlying computational ability. For addition, typical assumed primitives are recognizing a digit, writing a digit, and knowing the concepts of right, left, etc. The complete procedure network explicitly specifies all the subprocedures of a skill and can be evaluated or "executed," thereby simulating the skill for any given set of inputs. By itself, this network merely provides a computational machine that performs the skill and is not of particular import. However, the possible misconceptions in this skill are represented in the network by incorrect implementations associated with subprocedures in its decomposition called "bugs".[4] Each buggy version contains incorrect actions taken in place of the correct ones.[5] An extension to the network evaluator enables the switching in of a *buggy* version of a procedure that allows the network to simulate the behavior of that buggy subskill. This feature provides a computational method for determining the external behavior of the underlying bugs.

Inferring a Diagnostic Model of the Student

The problem of diagnosing a *deep structure* failure in a student's knowledge of a procedural skill can now be accomplished, at least theoretically, in a straightforward manner. Suppose, as in the examples on page 157, we are provided with several *surface* manifestations of a deep structure misconception, or bug, in the student's addition procedure. To uncover those possible subprocedures which are at fault, we use the network to simulate the behavior of buggy subprocedures over the set of problems and note those which generate the same behavior as exhibited by the student. To catch a student's misconceptions that involve more than one faulty subprocedure, we must be able to simulate various combinations of bugs. A student may have a bug in his carrying procedure as well as believing that 8+7 is 17 (a bug in his addition facts table). To model his behavior, *both* buggy versions must be used *together*. A *deep-structure model* of the student's errors is a set of buggy subprocedures that, when invoked, replicates those errors. Each buggy version has associated information such as what the underlying causes of the bug may have been, as well as specific remediations, explanations, interactions, and examples of the bug—all of which may be used by a tutoring system to help correct the student's problem.[6]

Many technical questions are raised by the above brief overview of how to "infer" a diagnostic model. We have deferred a more detailed discussion of these questions until Section 3 in favor of a more general discussion of pedagogical ramifications of the procedural network model of procedural skills. We begin this discussion with a description of the procedural network for a simple skill—subtraction.

---

[4] The term "bug" is borrowed from computer science where it refers to a mistake in a computer program.

[5] A "buggy" implementation does, in general, call other "correct" subprocedures though they may be called at inappropriate times or their results used incorrectly.

[6] West (1971) has broken down the diagnostic teaching task into four steps: (i) distinguish between conceptual and careless errors; (ii) identify the exact nature of the conceptual error (bug); (iii) determine the conceptual basis (cause) of the bug; and (iv) perform the appropriate remediation. The system we describe has been directed toward problems (i) and (ii). The buggy implementation nodes in the network provide the proper places to attach information relevant to problems (iii) and (iv).

A Procedural Network for Subtraction

As an example of the surprising amount of procedural knowledge needed to perform a simple skill, let us consider a more complete network representation of the subtraction of two numbers.[7] Figure 2 shows the links of the procedural network for subtraction that indicate which subprocedures a procedure may use.

The topmost node (SUBTRACT) represents the subtraction of two *n*-digit numbers. It may use the procedures for setting up the problem (SETUP), transforming it if the bottom number is greater than the top (TRANSFORM), and sequencing through each column performing the column subtraction (COLUMN SEQUENCE). The implementation of the column subtraction procedure has to account for cases where borrowing is necessary (BORROW NEEDED) and may call upon many other subprocedures, including taking the borrow from the correct place (DO BORROW), scratching 0 and writing 9 if that place contains a zero (ZERO), and so on. An important subprocedure is the facts table look-up (FACTS TABLE) which allows any of the simple arithmetic facts to be wrong. The facts table subprocedure is called during the addition of 10 to a column digit

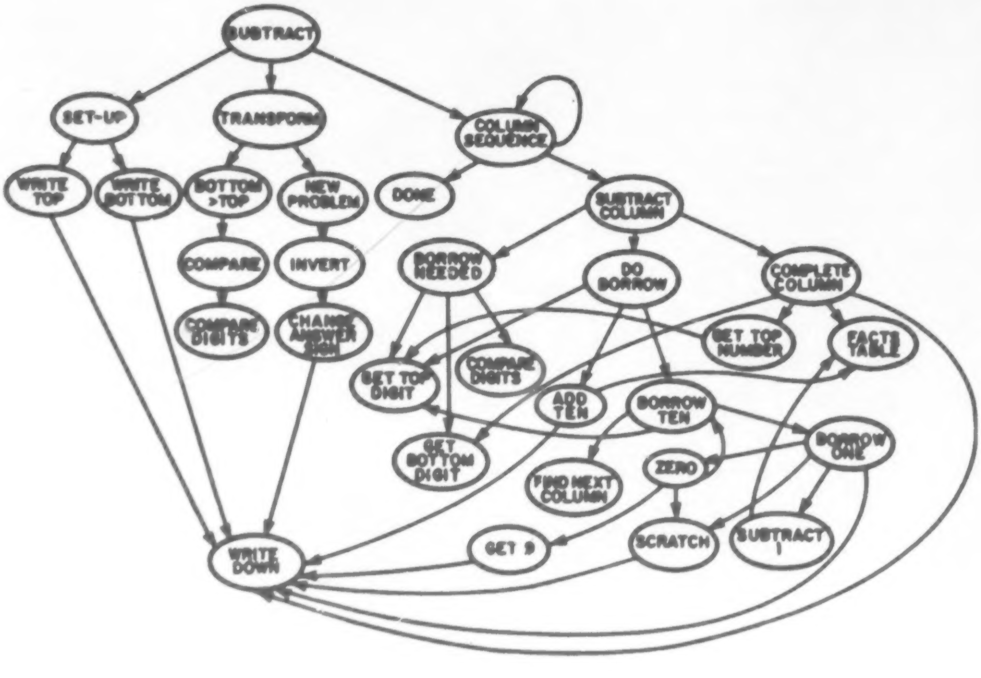

SUBTRACT
SET-UP
TRANSFORM
COLUMN SEQUENCE
WRITE TOP
WRITE BOTTOM
BOTTOM >TOP
NEW PROBLEM
DONE
SUBTRACT COLUMN
COMPARE
INVERT
BORROW NEEDED
DO BORROW
COMPLETE COLUMN
COMPARE DIGITS
CHANGE ANSWER SIGN
GET TOP NUMBER
FACTS TABLE
GET TOP DIGIT
COMPARE DIGITS
ADD TEN
BORROW TEN
BORROW ONE
GET BOTTOM DIGIT
FIND NEXT COLUMN
ZERO
GET 9
SCRATCH
SUBTRACT 1
WRITE DOWN

FIG. 2 A procedural network for subtraction.

---

[7] We have chosen just one of the several subtraction algorithms (the so-called "standard" algorithm), but the ideas presented here apply equally to others and can handle multiple methods as well.

$\begin{array}{r}143\\-28\\\hline 125\end{array}$	The student subtracts the smaller digit in each column from the larger digit regardless of which is on top.
$\begin{array}{r}143\\-28\\\hline 125\end{array}$	When the student needs to borrow, he adds 10 to the top digit of the current column without subtracting 1 from the next column to the left.
$\begin{array}{r}1300\\-522\\\hline 878\end{array}$	When borrowing from a column whose top digit is 0, the student writes 9 but does not continue borrowing from the column to the left of the 0.
$\begin{array}{r}140\\-21\\\hline 121\end{array}$	Whenever the top digit in a column is 0, the student writes the bottom digit in the answer; i.e., $0-N=N$.
$\begin{array}{r}140\\-21\\\hline 120\end{array}$	Whenever the top digit in a column is 0, the student writes 0 in the answer; i.e. $0-N=0$.
$\begin{array}{r}1300\\-522\\\hline 788\end{array}$	When borrowing from a column where the top digit is 0, the student borrows from the next column to the left correctly but writes 10 instead of 9 in this column.
$\begin{array}{r}321\\-89\\\hline 231\end{array}$	When borrowing into a column whose top digit is 1, the student gets 10 instead of 11.
$\begin{array}{r}662\\-357\\\hline 205\end{array}$	Once the student needs to borrow from a column, s/he continues to borrow from every column whether s/he needs to or not.
$\begin{array}{r}662\\-357\\\hline 115\end{array}$	The student always subtracts all borrows from the leftmost digit in the top number.

FIG. 3 Manifestations of some subtraction bugs.

(ADD 10), during the subtraction of 1 from a column digit in a borrowing operation (SUBTRACT 1), as well as during the subtraction of the bottom from the top digit in a column (COMPLETE COLUMN).

In principle, each of these subprocedures could have many buggy versions associated with it.[a] An example of a common bug is to calculate the column difference by subtracting the smaller digit from the larger regardless of which is on top. In another bug, the SETUP procedure left justifies the top and bottom numbers so that when the student is told to subtract 13 from 185, he gets 55. An interesting aspect of the left justification bug is that when the student is faced with seemingly impossible problems (185−75) he may be inclined to change the direction in which he subtracts, borrowing from right to left instead of from left to right, or to change his column difference procedure to larger minus smaller, thereby eliminating the need to borrow. Thus, there can exist relationships between bugs such that one bug suggests others. A major challenge in identifying the procedural breakdown or description of a skill is to have the network handle such ramifications and interactions of multiple bugs.

---

[a] In our current subtraction network, some subprocedures have only one buggy version, while others have as many as fifteen. The average is three or four.

To provide a feeling for the range of "answers" that can come from simple underlying bugs, we have included in Fig. 3 "answers" to subtraction problems using some of the bugs in the procedural network for subtraction. Notice that a particular answer to a given problem can have more than one explanation since several distinct bugs can generate the same answer. A special case is that a student may harbor many misconceptions and still get the correct answer to a particular problem!

The Power of Simulating Bugs in the Network

Given a procedural network like the one in Fig. 2, it is not always obvious how bugs in any particular subprocedure or set of subprocedures will be manifested on the surface, that is, in the answer. Some of the complicating factors are that a single buggy subprocedure can be used by several higher-order procedures in computing an answer or that two bugs can have interactions with each other. These factors are further complicated by the fact that not all sample problems will manifest all of the possible symptoms. If asked to make predictions about the symptoms of a given bug, people often determine the symptoms by considering only the skills or subprocedures used in solving *one* particular sample problem. As a result, they often miss symptoms generated by other procedures that can, in principle, use the given buggy subprocedure but which, because of the characteristics of the particular problem, were not called upon. If a different sample problem had been chosen, it might have caused the particular faulty subprocedure to have been used for a different purpose, thereby generating different symptoms. These observations first led us to consider the value of simulation to systematically verify a conjectured bug.

2. PEDAGOGICAL ISSUES

BUGGY: An Instructional Game for Training Student Teachers (and Others)

As we saw in Section 1, it is often difficult to infer a student's bug from his answers. It was the need for teachers to appreciate this problem and to develop strategies for coping with the possible range of student bugs that led us to construct a game called BUGGY. BUGGY is a computerized game based on the diagnostic interactions of a tutor and a student. The computer plays the part of an errant student whose malady must be identified by the diagnostician (a role played either by an individual or by a team). The diagnostician is presented with an arithmetic homework problem that the "student" has done incorrectly. The "student's" behavior is generated using the procedural network and shows the symptoms of an underlying bug in one of the arithmetic subprocedures. The diagnostician's job is to discover exactly what is the underlying misconception. He does this by providing strategic *test problems* for the "student" to solve. The problems given by the diagnostician are answered by the computer acting as the student using the bugged procedure.

The computer also acts as arbiter in the evaluation of the diagnostician's theory. When the diagnostician indicates that he has discovered the bug, he is asked to describe it. Then, to make certain that he really has found the bug, he is given a five-problem test during which he must answer the problems in the same errant way that the student would. In other words, he must simulate the "student's" bug. When the diagnostician's concept of the bug is sufficiently well formulated that he is able to predict the buggy answer on all five problems, he is considered to have diagnosed the student and he proceeds to a new bug.

We have found it beneficial to allow teachers to team up while playing the diagnostician's role in BUGGY. The interaction between team members encourages verbalization of the decision-making processes involved in the game.

The following is a protocol of a team of teachers using BUGGY. The dialogue is interspersed with commentary about the important issues. Those lines typed by the team are marked with a vertical line in the left margin.

Protocol of a Team Using BUGGY

```
WELCOME TO BUGGY.
I HAVE CHOSEN A BUG. HERE IS AN EXAMPLE OF THE BUG.

 17
 +5
 --
 13
```

The system presents an example in which the student does not get the right answer. At present the initial examples are stored in the network linked to the buggy implementation.

```
NOW YOU CAN GIVE ME PROBLEMS TO DETERMINE THE BUG.

|  18     43
|  +6    +79
   --    ---
   15     23
```

The team generates problems that are answered by the "student". Notice that the team's test problems are similar to the initial example generated by BUGGY. This is usually a poor diagnostic strategy as it may leave other symptoms of a bug undiscovered.

```
| Got the bug!
```

The team forms an initial hypothesis and tells the system they have discovered the bug. The system now asks the team to describe the bug and then tests them on their understanding. Asking for [illegible] description forces the team to verbalize their sometimes fuzzy idea of what the bug is and often makes them realize how ill-defined or vague their idea actually is. In the present system, the descriptions are saved but no attempt is made by the system to understand them.

PLEASE DESCRIBE THE BUG.

| Student adds the number to be carried into the last column worked in.

This cryptic description is typical of those we have observed. Describing bugs in English is surprisingly difficult. What the team is trying to say is that if a column adds to more than 9, the units digit and the carry digit are added together and the result is written as the answer. For example, in the first column of the example problem, when 7 and 5 are added, 1 (the carry) and 2 (the units digit) are added to get 3. The same with 8+6, 2+9, and 3+9. (This bug could easily arise if the student were using Method 2 from Fig. 1 and mistakenly wrote the carry under the same column.)

The hypothesis does account for the observed behavior. Unfortunately, it is not the student's bug. Notice the possibly humorous, probably disastrous results that would have transpired if the team had stopped at this point and tried to tutor the student on the basis of their initial belief!

HERE ARE SOME PROBLEMS TO TEST YOUR THEORY ABOUT THE BUG.
WHAT IS:

```
   21
  +39
  ---
|  51
```

To ensure that the team has found the correct bug, the system produces several problems for the team to solve. At present, BUGGY has a heuristic problem generation program that gives a broad range of "different" problems. (For examples in which the bug is very localized, such as requiring a 0 in the top number, a special problem generator can be linked to the bug and will be used instead.) These problems are then filtered to try to get three problems which cause symptoms and two that give correct answers. This simple routine does surprisingly well at producing critical problems. So well, in fact, that a more "intelligent" routine (which determined the set of bugs consistent with the student's data and then selected the test problems which split this set) was discarded as unnecessarily time-consuming. However, this more intelligent technique—based on techniques discussed in Section 3—might be reverted to as our theory of procedural bugs becomes progressively more subtle.

The team's hypothesis is revealed as incorrect by the first of the five test problems. BUGGY presents its answer and then allows the team to enter more problems.

THAT'S NOT THE BUG I HAVE.
USING MY BUG:

```
   21
  +39
  ---
   15
```

TRY GIVING ME SOME MORE EXAMPLES.

```
|    51      99     68
| +1707     +99     +9
   ----     ---     ---
     21      36     23
```

After failing the test, the team gives a greater range of problems, specifically those using bigger numbers. How is it that 51+1707 can add up to 21? By broadening the scope of example problems they are able to reevaluate their initial hypothesis, dismiss it, and develop another one that is complete and consistent with respect to the new errors.

| Got the Bug!

PLEASE DESCRIBE THE BUG YOU THINK I HAVE.

The student is adding all the digits together.

HERE ARE SOME PROBLEMS TO TEST YOUR THEORY ABOUT THE BUG.
WHAT IS:

```
 33   1091     8    28   90
+99    +60   +34   +70   +6
---   ----   ---   ---  ---
 24     17    15    17   15
```

This time the team's hypothesis passes the test and they are given the description of the bug stored in the procedural network.

VERY GOOD. MY DESCRIPTION OF THE BUG IS:
THE STUDENT ALWAYS SUMS UP ALL THE DIGITS WITH NO REGARD TO COLUMNS.

A variant of this game assigns a cost to each problem the user poses in the process of trying to discover the bug. At the same time, a much higher cost is assigned for claiming to have "discovered" the bug but then failing to pass criterion on the computer-generated test problems. A correctly structured cost algorithm encourages the user to think carefully about the diagnostic value of each problem he creates and not to jump to a conclusion too quickly or too slowly.

Pedagogical Issues for Teachers

One application of BUGGY and the "diagnostic model" view of procedural skills lies in the domain of teacher training. The realization that errors that appear "random" are often the surface manifestations of a systematic underlying bug is a major conceptual breakthrough for many student teachers. All too often, behavior that appears to be random has a simple, intelligent, and complete underlying explanation. By proper diagnosis, remediation for a student can be directed toward his specific weaknesses. The importance of simply admitting that there may exist underlying bugs cannot be overstressed. Without appreciation of this fact, a teacher must view failure on a particular problem as either carelessness or total algorithm failure. In the first case, the predicated remediation is giving more problems, while in the second, it is going over the entire algorithm.* When a student's bug (which may only manifest itself occasionally) is not recognized by the teacher, the teacher explains the errant behavior as carelessness, laziness, or worse, thereby often mistakenly lowering his opinions of the student's capabilities.

From the student's viewpoint, the situation is much worse. He is following what he believes to be *the* correct algorithm and, seemingly at random, gets marked *wrong*. This situation can be exacerbated by improper diagnosis. For example, Johnnie subtracts 284 from 437 and gets 253 as an answer. "Of course'', says the teacher ''you forgot to subtract 1 from 4 in the hundreds place when you borrowed.'' Unfortunately Johnnie's algorithm is to subtract the smaller digit in each column from the larger. Johnnie does not have any idea what the teacher is talking about (he never ''borrowed''!) and feels that he must be very stupid indeed not to understand. The teacher agrees with this assessment as none of his remediation has had any effect on Johnnie's performance.

---

*In computer programming metaphors, this approach corresponds to the debugging activities of resubmitting the program because the computer must have made a mistake and of throwing the whole program away and starting over from scratch and writing a new program.

BUGGY, in its present form, presents teachers with examples of buggy behavior and provides practice in diagnosing the underlying causes of errors. Using BUGGY, teachers gain experience in forming theories about the relationship between the symptoms of a bug and the underlying bug itself. This experience can be cultivated to make teachers aware that there are methods or strategies that they can use to diagnose bugs properly.

In fact, there are a number of *strategy bugs* that teachers may have in forming hypotheses about a student's misconceptions. That is, the task of diagnosing a student is a procedural skill and as such is susceptible to mislearning by the teachers. A common strategy bug is to jump too quickly to one hypothesis. Prematurely focusing on one hypothesis can cause a teacher to be unaware that there may be other competing hypotheses that are possibly more likely. A common psychological effect of this approach is that the teacher generates problems for the student that confirm his hypothesis! In some cases, teachers may believe their hypotheses so strongly that they will ignore contrary evidence or decide that it is merely random noise. One general diagnostic strategy that avoids this pitfall is the technique of differential diagnosis (Rubin, 1975) in which one always generates at least two hypotheses and then chooses test problems that separate them.

Another common strategy bug is to lock onto only one type of symptom. For example, one student teacher was given the initial example (A), after which he proceeded to generate example problems (B) and (C):

$$\begin{array}{ccc} \text{A} & \text{B} & \text{C} \\ 19 & 23 & 81 \\ \underline{+9} & \underline{+6} & \underline{+8} \\ 199 & 236 & 818 \end{array}$$

At this point, he concluded that the bug was ''writes the bottom digit after the top number.'' But his hypothesis failed when he was given the first test problem:

$$\begin{array}{r} 8 \\ \underline{+12} \end{array}$$

to which he responded 812. The bug is that single digit operands are linked on to the end of the other operand, so that the correct buggy answer is 128. By presenting examples only with a shorter bottom digit, he had obtained what seemed to be confirming evidence of his hypothesis. A general rule which could be employed to avoid this fixation is that whenever an example of incorrect behavior has an asymmetry (length of top and bottom numbers), then try an example with the asymmetry reversed. Using this rule, the teacher would also generate problems with larger top numbers before he reached a conclusion. BUGGY provides an environment in which teachers can experience the ramifications of not employing rules and strategies during diagnosis.[10]

Another important issue concerns the relationship between the language used to describe a student's errors and its effect on what a teacher should do to remediate it. Is the language able to convey to the student what he is doing wrong? Should we expect teachers to be able to use language as the tool for correcting the buggy algorithms of students? Or should we expect teachers only to be able to understand what the bug is and attempt remediation with the student with things like manipulative math tools? The following descriptions of hypotheses given by student teachers, taken from protocols of BUGGY, give a good idea of how difficult it is to express procedural ideas in English. The descriptions in parentheses are BUGGY's prestored explanations of the bugs.

> "Random errors in carryover." (Carries only when the next column in the top number is blank.)
>
> "If there are fewer digits on the top than on the bottom she adds columns diagonally." (When the top number has fewer digits than the bottom number, the numbers are left-justified and then added.)
>
> "Does not like zero in the bottom." (Zero subtracted from any number is zero.)
>
> "Child adds first two numbers correctly. Then when you need to carry in the second set of digits child adds numbers carried to bottom row, then adds third set of digits diagonally, finally carrying over extra digits." (The carry is written in the top number to the left of the column being carried from and is mistaken for another digit in the top number.)
>
> "Sum and carry all columns correctly until get to last column. Then takes furthest left digit in both columns and adds with digit of last carried amount. This is the sum." (When there are an unequal number of digits in the two numbers, the columns that have a blank are filled with the leftmost digit of that number.)

Even when one knows what the bug is in terms of being able to mimic it, how is one going to explain it to the student having problems? Considering the above examples, it is clear that anyone asked to solve a set of problems using these explanations would, no doubt, have real trouble. One can imagine a student's frustration when the teacher offers an explanation of why he is getting problems marked wrong, and the explanation is as confused and unclear as these are.

For that matter, when the correct procedure is described for the first time, could it, too, be coming across so unclearly!

The problem of adequately describing bugs is further complicated by another surprising fact: Fundamentally different bugs can cause identical behavior! In other words, there can be several distinct ways of incorrectly performing a skill that *always* generate the same "answers". For example, here is a set of problems:

---

[10] One job for an "intelligent" tutor for BUGGY is to recognize and point out places where instances of general rules and strategies should be used. This possibility is discussed in Section 4.

$$\begin{array}{r} 38 \\ +46 \\ \hline 174 \end{array} \qquad \begin{array}{r} 186 \\ +254 \\ \hline 2330 \end{array} \qquad \begin{array}{r} 298 \\ +169 \\ \hline 2357 \end{array} \qquad \begin{array}{r} 89 \\ +64 \\ \hline 243 \end{array}$$

One possible bug which accounts for these results is: the columns are added without carries and the leftmost digit in the answer is the total number of carries required in the problem. In this case, the student views the carries as tallies to be counted and added to the left of the answer. But another equally plausible bug also exists; the student places the carry to the left of the next digit in the top number; then, when adding that column, instead of adding the carry to the digit, he mistakes it as a tens column of the top digit—so that when adding 298 and 169, in the second column he adds 19 to 6 instead of 10 to 6. This generates the same symptoms. So even when the teacher is able to describe clearly what he believes is the underlying bug, he may be addressing the wrong one. The student may actually have either one of these bugs.[11]

An Experiment Using BUGGY with Student Teachers

To determine BUGGY's impact on student teachers, we had a group play the game described in the beginning of this section. The goal of the experiment was to explore whether exposure to BUGGY significantly improves the student teacher's ability to detect regular patterns of errors in simple arithmetic problems.[12] The subjects were undergraduate education majors from Lesley College in Cambridge. Their exposure to BUGGY lasted approximately one and a half hours, during which time both addition and subtraction bugs were presented. The effects of their exposure to BUGGY were measured by comparing each subject's performance on the pre- and postexposure debugging test. The detailed analysis and discussion of the experiment is beyond the scope of this paper but has been described in a technical report (Brown et al., 1977). Briefly, though, the results of the experiment showed that exposure to BUGGY significantly improved their ability to detect regular patterns of errors.

We also investigated the qualitative issue of what the student teachers felt they gained from their exposure to BUGGY. To assess their impressions, we convened the entire group after they had finished using BUGGY. At that gathering, we first asked them to write their responses to two questions (discussed below), and then we taped a final group discussion in which we sought their reactions to using

---

[11] This possibility leads to an interesting question concerning how one can "prove" that two different bugs entail logically the same surface manifestations.

[12] We would like to thank Dr. Mark Spikell of Lesley College for his assistance in this endeavor.

BUGGY and their suggestions for its deployment with school-aged students. The following week, their professor, who also participated in the initial experiment, held a second group discussion and reported to us the consensus, which was consistent with what they had written.

Appendix 1 lists some of the written responses to the question "What do you think you learned from this experience?". All 24 responded that they came away with something valuable. Many stated that they now appreciated the "complex and logical thought process" that children often use when doing an arithmetic problem incorrectly. "It makes me aware of problems that children have and they sometimes think logically, not carelessly as sometimes teachers think they do." "I never realized the many different ways a child could devise his own system to do a problem." They also stated that they learned better procedures for discovering the underlying bug. "I learned that it is necessary to try many different types of examples to be sure that a child really understands. Different types of difficulties arise with different problems."

We also asked the students "What is your reaction to BUGGY?" Many felt that "BUGGY could be used to sharpen a teacher's awareness of different difficulties with addition and subtraction." They felt that it might be of use in grade school, high school, or with special needs students, or even as a "great experience in beginning to play with computers."

Pedagogical Issues More Specific to Middle School Students

We feel that all of the issues discussed above are as important for school-level students as they are for teachers. There is great value in introducing young students to procedural notions. The BUGGY system provides a well-controlled environment for such an introduction, as well as one that can be *easily* integrated into a *standard curriculum*. Also note that for a middle school student to play the diagnostician's role requires his studying the procedural skill per se (i.e., its structure) as opposed to merely performing it. This experience can be especially important as students begin algebra, which is their first exposure to procedure "schema." By presenting procedures as objects of study, BUGGY can thus be used to explore the powerful ideas of hypothesis formation, debugging, debugging strategies, and so on. Of course, such a use requires more than just the BUGGY game.

Another reason for having students develop a language for talking about procedures, processes, bugs, and so forth is that such a language enables them to talk about, and think about, the underlying causes of their own errors. This facility is important in its own right, but it also gives a student the motivation and the apparatus for stepping back and critiquing his own thinking, as well as saying something interesting and useful about his errors. The difficulty in getting a student to test the plausibility of his own answer (such as by estimating it) may be due to the lack of any appropriate introspective skills that the student could use once he knew his answer was wrong.

An important ancillary, nonmathematical benefit of a student's involvement with BUGGY is exposure to the idea of role reversal.[13] To communicate effectively with others, children must learn not only language, but also the use of "social speech"—speech that takes into account the knowledge and perspective of another person (Krauss & Glucksberg, 1977). Piaget uses the term "childhood egocentrism" to describe the child's inability to detach himself from his own point of view and to take into consideration another's perspective. Although Krauss and Glucksberg agree that egocentrism plays a large part in very young children's speech, they believe that even in older children the ability to role-play breaks down when they are faced with a demanding cognitive task. We believe that taking on the viewpoint of the errant student by analyzing another's mistakes can provide valuable practice in role-playing in a demanding situation and can be beneficial to the development of "social speech."

Some Results on Using BUGGY with Seventh and Eighth Graders

To explore the effect of the BUGGY game on seventh and eighth graders and to discover what type of additional instructional material and activities must back up the use of this computer-based system, we placed a terminal running BUGGY in a classroom. The teacher provided a short introduction to the game and the notion of bugs and then the students[14] were free to use the system during the term. During this experience, we noticed the following phenomena. When the students first started trying to discover the underlying bugs in BUGGY, their most common reaction, upon seeing the mistakes of the simulated buggy student, was to exclaim "Wow, how dumb and stupid this kid must be." However, after a week or so of exposure to BUGGY, the students' reactions changed from believing that the simulated student was dumb to one of appreciating that there was, in fact, a systematic explanation for what this student was doing. They began to see that he (it) had fundamental misconceptions as opposed to being just stupid. This result is particularly exciting, since it paves the way for students to see their own faulty behavior not as being a sign of their stupidity, but as a source of data from which they can understand their own errors.

Unfortunately, we have as yet no data concerning the transferability of this awareness, and whether or not it will lead to students being more capable and willing to look over their own work for errors.

---

[13] This idea is attributable to Tim Barclay, at the Cambridge Friends School, who has been experimenting with various uses of BUGGY with sixth through eighth graders.

[14] All of the students participating in this activity had already mastered the procedural skills being "bugged." In fact, we have severe reservations about the advisability of younger students using this particular game especially if they have not mastered the correct versions of the skills. However, there are extensions to BUGGY (discussed in Section 4) that make it appropriate for school children in the process of learning the given procedural skills.

3. AN AUTOMATED DIAGNOSTIC MODELING SYSTEM

In this section, we describe a diagnostic system that is based on our arithmetic procedural network, recently completed by Burton. The notion of modeling procedural skills as *executable* networks and then expressing all potential misconceptions as buggy versions of subprocedures in the network provides a technique for efficiently determining the consequences or symptoms of a given collection of bugs on a set of problems. It therefore has the potential of diagnosing and explaining all of the procedurally incorrect answers for any given problem. For example, as we indicated in Section 1, given a procedural network for addition and an addition problem (like 35+782), all the buggy subprocedures as well as combinations of buggy subprocedures, can be inserted and then executed in the addition procedure one by one so that all the possible buggy answers to the problem are generated. Those answers can then be compared to a student's work to determine possible explanations of the student's particular misconceptions. In this way, one might use the procedural network to diagnose a student's basic misconceptions over an entire homework assignment or an arithmetic test.

There are, however, several complications with this simple paradigm for diagnosing a student. One is that the student who has developed a *novel* singleton bug (as opposed to one that arose out of a combination of "primitive" bugs) will not be diagnosed. Another is that students *do* make "random" mistakes (presumably as many while following an *incorrect* procedure as a correct one) that could erroneously lead to the exclusion of his bug or the inclusion of another bug that happened to coincide with his "randomness." Finally, blindly considering all possible combinations of bugs can lead to a combinatorial explosion of possibilities. In what follows, we shall discuss some solutions to these problems.

A Deep-Structure Data Analysis Tool

As a first step in developing techniques for automatically diagnosing a student's errors, we sought a large data base of student answers on some arithmetic test. We started out this way for two reasons. First, an analysis of the student errors not already explained by the existing network can suggest extensions to the network that may have been overlooked. Obviously, if the BUGGY system is going to be useful as a diagnostic tool, its arsenal of bugs must be very extensive! Second, once a "complete" network of bugs has been constructed, analyzing a large data base can provide some evidence of how many student errors are procedural rather than "careless" errors. Such an analysis also indicates how consistently students apply buggy procedures. Do they use the buggy procedure every time it is appropriate? When they get a correct answer does that contraindicate their use of a buggy procedure, or is it just that the buggy procedure for this problem *does* produce the correct answer? This analysis also reveals which procedural errors occur most often, and in what combinations. Answers to these questions not only influence the design of the diagnostic system (as we will see later) but also have an impact on how that system could be used.

We were fortunate to be able to obtain a large collection of pertinent data[15] already in computer readable form. This data stemmed from an achievement test administered in Nicaragua to fourth, fifth, and sixth graders. There were 10 different test versions, each consisting of 30 problems combining both simple and complex addition and subtraction problems. One version of the test is given in Fig. 4. The makeup of each test followed a complex procedure discussed by Friend (1976).

1 +2	8 -3	7 +9	99 -79	43 +41	353 -342	213 21	633 -221	521 502	81 -17
123 13 610 +12	4769 -0	9 91 +6	257 -161	597 +75	6523 -1280	156 873 +311	103 -64	8 54 +9	7315 -6536
505 743 12 +35	1039 -44	77 18 +47	705 -9	917 639 +5	10038 -4319	864 9 4 +3	10060 -98	579 96 833 +43	7001 -94

FIG. 4 A sample test.

Admittedly, these data have the limitation that the particular results derived from them are not necessarily generalizable to American schools. Although the procedures taught for addition and subtraction are similar, the environmental and cultural experiences of the students are quite different. Nevertheless, this data base provides a convenient starting point for this research as well as a general idea of the percentage of students making procedural as opposed to "random" errors. In addition, the methods we devised to analyze these data would apply equally well to data collected under other circumstances.

The Process of Organizing the Data

The data base available for this study was large—19,500 problems performed by 1300 students. We limited ourselves to consideration of just the subtraction problems on the test because the addition problems included some in which three

---

[15] The data used in this study were made available by the Institute for Mathematical Studies in the Social Sciences at Stanford University. The data were collected in Nicaragua as part of the Nicaragua Radio Mathematics Project supported by Contract A10/CM-ta-C-73-40 from U.S. Agency of International Development. We are grateful to Barbara Searle for allowing us access to the data. See Searle, Friend, and Suppes (1976) for a description of the NRM project.

TABLE 1
An Initial Bug Comparison Table[a]

	8	99	353	633	81	4769	257	6523	103	7315	1039	705	10038	10060	7001
	3	79	342	221	17	0	161	1280	64	6536	44	9	4319	98	94
	5	20	11	412	64	4769	96	5243	39	779	995	696	5719	9962	6907
Student answers	—	—	—	—	98	—	418	—	169	738	1095	706	14319	10078	7097
FORGET/BORROW/OVER/BLANKS	*	*	*	*	!	*	!	*	139	!	***	***	15719	10062	7007
STOPS/BORROW/AT/ZERO	*	*	*	*	!	*	!	*	49	!	***	***	6719	10062	7017
DIFF/$0-N=N$	*	*	*	*	!	*	!	*	!	839	!	!	***	9978	!
ADD/INSTEADOF/SUB	11	178	695	854	***	*	***	7803	167	13851	1083	714	14357	10158	7095

[a] This table presents how well a student's answers are explained by different bugs; "***" and "*" indicate places of agreement. (See text for discussion.)

or more numbers were to be added—a condition that can produce errors not modeled in the present addition network.

As a first step, we extended the BUGGY system to print out "bug comparison" tables for students as shown in Table 1. These tables summarize how well the student's behavior can be explained or predicted by a simple bug in the network. The format of a table is as follows: The problems with the correct answers appear at the top of each table. The student's answers appear on the next line using the convention that "—" indicates a correct student answer. Each of the remaining lines provides the name of a bug and the answers produced by the assumption that the student had this and only this bug. For each of these lines a "***" means that the bug predicted the student's incorrect answer. A "*" means that the bug in that row would give the correct answer and also that the student got the correct answer. Thus "*" and "***" indicate places of agreement between the student's behavior and the simulated behavior of the bug. An "!" means that the bug would give the correct answer, but that the student gave an incorrect one.[16] A number which appears in a bug row is the answer that the bug would give when it is different from both the student's answer and the correct answer.[17]

---

[16] In both the "*" and the "!" cases, the bug has not manifested itself in the answer as an error. For example, if a bug is $0-N=0$, it would not show itself in a problem unless there is a 0 in the problem's top number (or 0 is generated during solution by borrowing from a column with a 1).

[17] An additional case arises when the student does not answer a problem. Although none of our example tables will include this case, it is marked with a "#". "#" is also used in a bug row to indicate that the bug could not do the problem. We saw very little evidence of students not doing a problem possibly because the students were given as much time as they wanted to complete the tests.

Initially, any bug that explained any of the student's behavior (i.e., generated at least one "***") was included in the table. However, these tables proved to be too large to conveniently read so a routine was added to delete any bug if there was another bug that accounted for the same set of answers as well as some others.

Analysis of a Bug Comparison Table

When we are determining whether or not a student has a particular bug, "*" and "***" are confirming evidence that the student has the bug while both "!" and numbers in the bug row are disconfirming evidence. We refer to the results in a bug comparison table as "evidence" because there may be several possible explanations for any particular answer to one problem. The student may have made a careless error while following his bugged procedure, therefore leading to a number in that bug row instead of a "***". He may have an unmodeled combination of bugs, only one of which manifested itself and resulted in an "!" in the row of the other bug. Or he may have been following a totally different procedure, or no procedure at all, that just happened to give him the same answer as a bug leading to a "***" in a bug row. The final decision on whether or not the student is using a buggy subprocedure must be made by considering all of the evidence from the test. But how should *conflicting evidence* be weighted and summed?

Let us consider an analysis of the bug comparison table for a particular student, Table 1. Both FORGET/BORROW/OVER/BLANKS and STOPS/BORROW/AT/ZERO produce the same answer in problems 11 (1039−44) and 12 (705 - 9)—the errant student answers. But neither has particularly good agreement across the rest of the table. Which, if either, misconception was the student operating under? Our inclination is to believe that neither bug satisfactorily explains the behavior, but how does one decide in general? To answer these questions, we analyzed, by hand, several hundred students' tables like the one in Table 1. During these formative analyses, the tables were examined to ferret out students whose behavior was not captured by any existing bugs. The work of these students was closely scrutinized for any underlying computational pattern. If a pattern could be discerned, the incorrect subprocedure was defined and this new "bug" was added to the network. During this formulation period our list of bugs grew from 18 to 60.

Multiple Bugs and Their (Nonobvious) Interactions

Most of the 60 subtraction bugs discovered during this period were primitive, in the sense that each redefined only one subprocedure in the subtraction network. Was it possible that some of the students' behavior was due to multiple bugs that we had failed to notice? To explore this possibility, we programmed the BUGGY diagnostic system to try all *pairs* of bugs. That is, buggy definitions of

TABLE 2
Multiple Bug Comparison Table

8	99	353	633	81	4769	257	6523	103	7315	1039	705	10038	10060	7001
3	79	342	221	17	0	161	1280	64	6536	44	9	4319	98	94
5	20	11	412	64	4769	96	5243	39	779	995	696	5719	9962	6907
Student answers														
—	—	—	—	98	—	418	—	169	738	1095	706	14319	10078	7097
DIFF/0−*N*=*N* and STOPS/BORROW/AT/ZERO														
*	*	*	*	!	*	!	*	***	839	***	***	***	***	***
ADD/INSTEADOF/SUB														
11	178	695	854	***	*	***	7803	167	13851	1083	714	14357	10158	7095

two subprocedures were systematically inserted and then executed.[18] This process turned up 270 bug combinations whose symptoms were different from any of the primitive bugs and from each other in one test of 15 problems.

In order to illustrate the diagnostic power of this generative technique, consider the example of the student whose work is shown in Table 1. From Table 1, no discernible pattern is evident. The student's work does however admit to a beautifully simple characterization which is the composite of two primitive bugs. Table 2 shows the new bug comparison table that was generated for this student from comparisons using multiple bugs. Notice that the comparison line which resulted from the combination of the two bugs is substantially different from either of the single bug lines or even from a linear combination of the two comparisons. This is due to the nonobvious interactions of the two bugs, particularly where the intermediate products of one of the bugs enables or disables the other bug. For example, in problem 9 (103–64), the "0−*N*=N" bug alone will not manifest itself because the borrow from the first column will have changed the 0 in the second column to 9. However, the "stops-borrow-at-zero" bug[19] has the side effect of *not* changing the 0 to 9, and hence enables the "0−*N*=0" bug. In general, the interactions between bugs can be arbitrarily complex. This can make a teacher's diagnostic task very difficult.

---

[18] During this process, bugs which were alternative definitions of the same conceptual procedure, of course, could not be paired. There were also cases where one bug would preclude another. For example, SMALLER/FROM/LARGER precludes any of the bugs in the borrowing procedures as borrowing is never required. In these cases, a bug can prevent other portions of the network from ever being executed and hence "switching in" bugs in the unused portion would be useless as long as the higher bug remained in effect. Rather than an extensive analysis of the potential interactions of bugs, we opted for the simpler solution of comparing the bugs via their symptoms over a fixed set of problems.

[19] In the "stops-borrow-at-zero" bug, the student does not know how to borrow from a column which has a 0 in the minuend so he does not do anything to it. He does however add 10 to the column he is processing.

Judging Diagnostic Credibility

Using the multiple bug comparison tables generated for about one hundred students, we identified which of the students, in our judgment, were making procedural, as opposed to careless or random errors. During this hand-done classification process, we articulated and refined our intuitive use of the evidence from a student's entire test in order to make that decision. Eventually, our understanding and description of the process became precise enough to be computerized so that it could be run on all 1300 students.

Our hand-done study suggested six intuitive groupings of students:

1. Those students who got all the problems correct.
2. Those students who erred on any number of problems but whose errors were explained by one bug or one bug pair.
3. Those students who clearly exhibited the presence of a bug but who also exhibited some behavior that was not explained by the bug.
4. Those students who missed only one or two (of fifteen problems) and in a way not consistent with any bug.
5. Those students who exhibited some buggy behavior but not consistently.
6. Those students whose behavior appeared random relative to the known bugs.

Tables 3, 4, and 5 show representative students from Groups 2, 3, and 5. The intuitive justification for these groupings stemmed from the possible tutorial approaches a teacher might take to remediate a student. The classes of students and the possible very general tutorial approaches we saw are:

- i. those students who are correct or very nearly correct and probably just need more practice if anything (Groups 1 and 4);
- ii. those students who are exhibiting consistently incorrect behavior and, therefore, whose remediation may profitably be viewed as a process of "debugging" the student's present algorithm (Groups 2 and 3), and

TABLE 3

Example of a Student Whose Behavior is Well Explained by One Bug[a]

8	99	353	633	81	4769	257	6523	103	7315	1039	705	10038	10060	7001
3	79	342	221	17	0	161	1280	64	6536	44	9	4319	98	94
5	20	11	412	64	4769	96	5243	39	779	995	696	5719	9962	6907
Student answers														
—	—	—	—	—	—	—	—	139	—	1995	76	15719	10962	7007
BORROW/FROM/ZERO														
*	*	*	*	*	*	*	*	***	*	***	796	***	***	***
Best guess: BORROW/FROM/ZERO														
GROUP=2														

[a] The bug, BORROW/FROM/ZERO, is that when borrowing from a column in which the top number is 0, the student writes 9 but does not continue borrowing from the next column to the left.

TABLE 4
Example of a Student Who Exhibits a Consistent Bug but Also Has Other Problems

	8	99	353	633	81	4769	257	6523	103	7315	1039	705	10038	10060	7001
	3	79	342	221	17	0	161	1280	64	6536	44	9	4319	98	94
	5	20	11	412	64	4769	96	5243	39	779	995	696	5719	9962	6907
Student answers	—	—	—	—	74	—	—	—	9	1209	95	704	10019	70	6007
DIFF/0−*N*=0 and MOVE/OVER/ZERO/BORROW	*	*	*	*	!	*	*	*	***	809	***	606	***	***	***
BORROW/ACROSS/SMALLER/ADDING/TEN/EXCEPT/ZERO and DIFF/0−*N*=0	*	*	*	*	!	*	*	*	***	889	***	606	***	***	***
SUB/UNITS/SPECIAL and SMALLER/FROM/LARGER&0−*N*=0	*	*	*	*	!	*	116	5363	!	***	1015	!	***	10042	!
BORROW/NO/DECREMENT and QUIT/WHEN/BOTTOM/BLANK	*	*	*	*	***	9	196	5343	49	1889	***	6	6729	72	17
BORROW/NO/DECREMENT	*	*	*	*	***	*	196	5343	149	1889	1095	706	16729	10072	7017
SMALLER/FROM/LARGER	*	*	*	*	76	*	116	5363	161	1221	1015	***	14321	10038	7093

Best guess: DIFF/0−*N*=0 and MOVE/OVER/ZERO/BORROW
GROUP=3

iii. those students for whom a thorough reteaching of the entire algorithm appears to be essential (Groups 5 and 6).

We are definitely *not* saying that for all students in Groups 2 and 3 the only or even the best pedagogy is to focus on the student's buggy procedure. Instead, we are trying to identify those students who are consistently making the same mistake and for whom debugging of their procedures *may* be useful.

TABLE 5
Example of a Student Who Exhibits Some Buggy Behavior but not Consistently

	8	99	353	633	81	4769	257	6523	103	7315	1039	705	10038	10060	7001
	3	79	342	221	17	0	161	1280	64	6536	44	9	431	98	94
	5	20	11	412	64	4769	96	5243	39	779	995	696	5719	9962	6907
Student answers	!	!	!	!	70	!	98	!	109	839	1095	706	1431	10700	7001
DIFF/0−*N*=*N* and STOPS/BORROW/AT/ZERO	*	*	*	*	!	*	!	*	169	***	***	***	***	10078	7097
DIFF/0−*N*=*N* and FORGET/BORROW/OVER/BLANKS	*	*	*	*	!	*	!	*	139	***	***	***	***	1078	7007
DIFF/0−*N*=0&1−*N*=1 and STOPS/BORROW/AT/ZERO	*	*	*	*	71	*	!	*	***	809	***	***	10019	10070	***
ZERO/INSTEADOF/BORROW	*	*	*	*	***	*	106	5303	100	1000	1005	700	10020	10000	7000

GROUP=5

Classification Algorithm

The algorithm that we eventually converged on for assigning a student to one of the six major categories defined in the previous section is rather involved, so some readers may prefer to skim the following discussion. If the student erred on at least one problem, each bug was rated according to how well it accounted for this behavior. This rating results in a group number for each bug.[20] The *rating* of each bug depends on the number of answers falling into each of five groups:

1. those student answers for which the bug predicts the student's incorrect answer (the number of "***"s appearing in the bug's line referred to as $N***$);
2. those student answers for which the bug predicts the student's correct answer (the number of "*"s—$N*$);
3. those student answers for which the bug predicts the correct answer but which the student answered incorrectly (the number of "!"s—$N!$);
4. those student answers for which the bug predicts an incorrect answer but which the student answered correctly (Nr);
5. those student answers for which the bug predicts an incorrect answer different from the student's incorrect answer (Nw).

This analysis gives a symptom vector of five numbers ($N***, N*, N!, Nr, Nw$).

From the symptom vector, a group number corresponding to the six major categories of student behavior given in the previous section is calculated for *each bug* using the following procedure:[21]

A bug indicates Group 2 student behavior if it agrees with all of the student's answers ($N! + Nr + Nw = 0$) or if it agrees more than 75% of the time on problems in which it predicts a wrong answer [$N*** \geq 3 \times (N! + Nr + Nw)$].

A bug indicates Group 3 behavior if it explains two or more student errors and predicts more correct than incorrect answers ($N*** > Nr + Nw + N!/2$). In this formula, those problems in which the bug did not exhibit a symptom are weighted by half—the intuition is that, on these problems, the student may be exhibiting other bugs as well.

A bug is also rated as indicating Group 3 behavior if it is a primitive bug (not multiple) that predicts more than half of the student's errors, and predicts erroneous behavior more times than it fails to do so ($N*** > N! + Nw$ and $N*** > Nr$).

A bug indicates Group 5 behavior if it predicts at least two incorrect answers ($N*** \geq 2$).

Otherwise a bug is rated Group 6.

---

[20] In the actual implementation, the bugs are ordered by the number of symptoms accounted for and are rated from the most promising bug until the group number increases.

[21] As we have said, the classification scheme was based primarily on empirical studies. There are intuitive justifications (rationalizations) for each of the decisions; however, in the final analysis, this algorithm was used because it classified students in close accordance with our hand-done analysis.

TABLE 6
Totals and Percentages of Student Classifications[a]

	Group													
	1		2		3		4		5		6		Totals	
Grade	No.	(%)	No.	(%)	No	(%)	No.	(%)	No.	(%)	No.	(%)	No.	(%)
4th	10	2.4	101	20.5	93	18.9	31	6.6	197	39.5	72	14.7	504	100.0
5th	10	3.0	86	22.0	73	18.7	38	10.0	132	33.5	60	15.5	399	100.0
6th	17	4.5	88	21.3	64	15.6	47	11.6	131	31.5	75	18.2	422	100.0
Totals	37	3.2	275	21.2	230	17.8	116	9.2	460	35.2	207	16.1		

[a] Groups 2 and 3 are consistently following an incorrect procedure (see text for further explanation of groups).

The student is assigned the group number of the lowest rated bug. If the lowest rated bug is not Group 2 or 3 and the student has missed only one or two problems, he is put in Group 4. If the student is put in Group 2 or 3, the bug with the lowest group rating (and which accounts for the most symptoms in cases of ties) is chosen as the most likely hypothetical student bug. Examples of this result can be seen in the last line of Tables 3 and 4.

Diagnostic Results for the Nicaraguan Data Base

The above classification procedure was used to analyze the set of test responses for 1325 fourth, fifth, and sixth graders.[22] A summary of the diagnostic classification by grade is given in Table 6. As can be seen, nearly 40% of the students exhibited consistently buggy behavior. This figure agrees with a similar result reported by Cox (1975). The similarity across grade level may be due to the fact that addition and subtraction are not presented again after the 4th grade.

Table 7 gives the frequency of the 14 most common bugs. Most of the difficulties arise while borrowing, especially when a zero is involved. The most common bug was "when borrowing from a column in which the top digit is 0, change the 0 to a 9 but do not continue borrowing from the next column to the left"; it occurred alone or together with other bugs 153 times in the 1325 student tests.

What Does a Test Score Mean?

One of the ramifications of this fully automatic diagnostic technique concerns its ability to score tests based on what a student knows or does not know as

---

[22] An implementation note: The entire data analysis program including the BUGGY subtraction models is written in INTERLISP. The analysis of all 1325 students against the 330 bugs required on the order of 90 minutes of CPU time on a PDP-KL10.

TABLE 7
Bug Frequency Table

The 14 most frequently occurring bugs in a group of 1325 students

57	students used: BORROW/FROM/ZERO (103−45=158) When borrowing from a column whose top digit is 0, the student writes 9, but does not continue borrowing from the column to the left of the 0.
54	students used: SMALLER/FROM/LARGER (253−118=145) The student subtracts the smaller digit in a column from the larger digit regardless of which one is on top.
50	students used: BORROW/FROM/ZERO and LEFT/TEN/OK (803−508=395) The student changes 0 to 9 without further borrowing unless the 0 is part of a 10 in the left part of the top number.
34	students used: DIFF/$0-N=N$ and MOVE/OVER/ZERO/BORROW Whenever the top digit in a column is 0, the student writes the bottom digit in the answer; i.e., $0-N=N$. When the student needs to borrow from a column whose top digit is 0, he skips that column and borrows from the next one.
14	students used: DIFF/$0-N=N$ and STOPS/BORROW/AT/ZERO Whenever the top digit in a column is 0, the student writes the bottom digit in the answer; i.e., $0-N=N$. The student borrows from zero incorrectly. He does not subtract 1 from the 0 although he adds 10 correctly to the top digit of the current column.
13	students used: SMALLER/FROM/LARGER and $0-N=0$ (203−98=205) The student subtracts the smaller digit in each column from the larger digit regardless of which one is on top. The exception is that when the top digit is 0, a 0 is written as the answer for that column; i.e., $0-N=0$.
12	students used: DIFF/$0-N=0$ and MOVE/OVER/ZERO/BORROW Whenever the top digit in a column is 0, the student writes 0 in the answer; i.e., $0-N=0$. When the student needs to borrow from a column whose top digit is 0, he skips that column and borrows from the next one.
11	students used: BORROW/FROM/ZERO and DIFF/$N-0=0$ When borrowing from a column whose top digit is 0, the student writes 9, but does not continue borrowing from the column to the left of the 0. Whenever the bottom digit in a column is 0, the student writes 0 in the answer; i.e., $N-0=0$.
10	students used: DIFF/$0-N=0$ and $N-0=0$ (302−192=290) The student writes 0 in the answer when either the top or the bottom digit is 0.
10	students used: BORROW/FROM/ZERO and DIFF/$0-N=N$ When borrowing from a column whose top digit is 0, the student writes 9, but does not continue borrowing from the column to the left of the 0. Whenever the top digit in a column is 0, the student writes the bottom digit in the answer; i.e., $0-N=N$.
10	students used: MOVE/OVER/ZERO/BORROW (304−75=139) When the student needs to borrow from a column whose top digit is 0, he skips that column and borrows from the next one.
10	students used: DIFF/$N-0=0$ (403−208=105) Whenever the bottom digit in a column is 0, the student writes 0 in the answer; i.e., $N-0=0$.
10	students used: DIFF/$0-N=N$ (140−21=121) Whenever the top digit in a column is 0, the student writes the bottom digit in the answer; i.e., $0-N=N$.
9	students used: DIFF/$0-N=N$ and LEFT/TEN/OK (908−395=693) When there is a 0 on top, the student writes the bottom digit in the answer. The exception is when the 0 is part of 10 in the left columns of the top number.

opposed to scoring it based *solely* on the number of right and wrong answers on his test. Even with the advances in criterion referenced testing, it remains true that a test is simply scored by what problems a student gets right or wrong. Because of the embedded nature of most procedural skills, a student can get many problems wrong simply by having one fundamental underlying bug in a primitive subprocedure that he is using to solve different problems on the test. In such a situation, the score that a student gets can bear little relationship to the misconceptions that he actually harbors! A student can receive a low score either because he has many misconceptions, each one of which is more or less the top of the procedural network of skills he is using, or he may possess a few, or even just one, misconception that is deep down inside the internal workings of the procedural network and which is constantly being used to compute intermediate results used by higher-up subprocedures.

Current techniques for correcting tests do not offer easy and reliable methods of separating these two situations. The diagnostic modeling technique discussed above can take the answers that a student gives on a test and, through its modeling system, show not only which questions were answered incorrectly but *why* they were incorrectly answered. It is interesting to note that 107 of the 1325 students tested had a bug in their borrow-from-zero subprocedure and missed 6 out of the 15 problems on the test because of this one underlying bug. The characterization given by BUGGY is a much fairer evaluation than scoring these students 60% correct.

A Methodological Tool for Judging the Diagnostic Quality of a Test

The procedural network apparatus also provides a methodological tool for judging the quality or diagnostic capabilities of a given test. This allows one to talk about how well the test can discover and delineate common misconceptions. Given a test, each bug in the network can be used to answer all of the problems. The resulting "buggy" responses are then used to partition the bugs: two bugs are put in the same partition if they produce the same answers to all of the test problems. "Diagnosticity" of the test can now be defined in terms of the size of the resultant partitions. A diagnostically perfect test has every bug in a partition by itself. Those bugs in the same partition are undifferentiated by the test. Any bugs in the same partition as the correct answers are not tested for at all. Taking the test in Fig. 4, this system discovered that a student can have either of two bugs (BORROW/ACROSS/SAME or $N-N=1$/AFTER/BORROW) and still get 100% correct answers.

Such a system could provide professional test designers with a formal tool for establishing the diagnostic quality of a proposed test. However, our belief is that professional test designers have good intuitions about diagnostic tests. This belief was confirmed by running two standardized national tests through the 330 subtraction bugs. One of the tests left only one bug unexposed in 17 problems while the other left 4 bugs unexposed in 10 problems. The unexposed bugs were rare for they were not even found in the Nicaraguan data. There is a difference, however, between exposing a bug and diagnosing it. *Exposing* a bug amounts to having at least one problem on the test in which the bug is manifested; *diagnosing* a bug requires having test problems which differentiate between it and every other bug.

Using the Artificial Intelligence paradigm of generate-and-test, it would be straightforward to use BUGGY as a diagnostic test generator. The problem generator must produce "interestingly" different problems. This generation can be done using important features of problems such as the number of times borrowing is necessary, or whether or not a zero appears in the top number. Sets of generated problems can then be filtered using the procedural network to identify bugs which are not diagnosed. From the bugs left undiagnosed, features can be retrieved which direct the generation of alternative problems to be added to the test. In this way, a highly diagnostic test can be developed. Furthermore, since the answers that would be generated by using the bugs are known, the test could be a multiple choice test and still maintain its total diagnostic property! Similarly, a real-time adaptive testing system could be created based on these tools.

4. FUTURE RESEARCH

This paper has presented some of the problems that must be faced in diagnosing failures in procedural skills, and has described some ideas about the formulation and implementation of diagnostic modeling techniques that address these problems. It has also presented some novel uses of diagnostic models as a gaming/instructional device, as a deep-structure test grader, and as a tool to judge the diagnostic quality of a test or a set of problems. The central idea underlying this research is the use of a *procedural network* as a means of representing diagnostic models. The critical properties of this representation scheme are its abilities to represent an appropriately structured breakdown of a skill into subskills, to make explicit the control structures underlying a collection of skills, and to make the knowledge encoded this way directly executable. Such a representation enables a particular subskill to be easily modified and then simulated or executed so that the ramifications of the modification can be quickly ascertained. The structure of the network becomes important not only because it allows efficient modification but also because the representation of the modification can be used to contain explanatory or remedial material.[23] In addition, the structure allows certain types of control structure failures to be directly represented in the network and hence articulated—if necessary.

---

[23] Contrast this with the (admittedly strawman) technique of randomly switching instructions in a machine language program which carries out a skill. Even if a student's behavior could be duplicated, the resulting "model" would be worthless as an explanatory device or as an aid to remediation.

Relationship of Diagnostic Models to Other Kinds of Structural Models

We now turn to a brief look at the past and current work on structural models of students and how they relate to diagnostic models based on procedural networks. Most of the past and current research on this subject has been focused on the intuitively appealing notion that if one has an explicit, well formulated model of the knowledge base of an expert for a given set of skills or problem domain, then one can model a particular student's knowledge as a simplification of the rules comprising the expert's procedures (Barr, 1974; Brown, 1974; Brown & Burton, 1975; Burton & Brown, 1977; Carbonell & Collins, 1973; Carr & Goldstein, 1977; Collins, Warnock, & Passafiume, 1975). Recently, Goldstein has expanded this concept in his Computer Coach research and has coined the term "overlay model" for capturing how a student's manifested knowledge of skills relates to an expert's knowledge base (Goldstein, 1977).

The work reported in this paper differs in that the basic modeling technique is based on viewing a structural model of the student not as a simplification of the expert's rules but rather as a set of *semantically meaningful deviations* from an expert's knowledge base.[24] Each subskill of the expert is explicitly encoded, along with a set of potential misconceptions of that subskill. The task of inferring a diagnostic model then becomes one of discovering which set of variations or deviations best explains the surface behavior of the student. This view is in concert with, although more structured than, the approach taken by Self (1974) in which he models the student as a set of modified procedures taken from a procedural expert problem-solver.

Another closely related approach to modeling a student's knowledge base uses a production rule encoding scheme (Smith & Sleeman, 1977; Young, 1977). However, procedural networks differ both theoretically and computationally from these efforts in that they are designed to make explicit the representation of the control-structure knowledge underlying a macro skill so that it can be efficiently diagnosed and explicitly tutored.

In the remainder of this section, we present our view of the more promising directions for research relating to diagnostic models.

Extensions to the Gaming Environment

In the second section, we described the BUGGY game which was designed to introduce the notion of "buggy" behavior and provide practice in diagnosing it. Although this activity was initially designed for training student teachers in diagnosing and articulating procedural bugs, it has also been used as an activity to get kids to introspect on their already known procedures as well as to encounter the concept of bugs and debugging strategies in an easily grasped context.

---

[24] Because these deviations are based on both the student's intended goals and the underlying teleology of the subskills, we have no automatic way to generate them (as opposed to what could be done if the deviations were based on the surface syntax of the rules). However, ongoing work by the authors, as well as Miller and Goldstein (1976) and Rich and Schrobe (1976), is directed toward helping to overcome this limitation.

A limitation of the current gaming environment is that most of what the players learn while using the game they learn or discover on their own. At the moment, BUGGY does no explicit tutoring; it simply provides an environment that challenges their theories and encourages them to articulate their thoughts.[25] The rest of the learning experience occurs either through the sociology of team learning or from what a person abstracts on his own. The next step in realizing the educational potential of the BUGGY game is to implement an intelligent tutor which can recognize and point out weaknesses (or interesting facets) in a student's debugging strategies.[26] Our experiences indicate that such a tutor would be very helpful for middle school and remedial students who often get caught in unproductive ruts. The tutor could also help focus the student's attention on the structure of the arithmetic procedures themselves. It is worth noting that *some* of the tools for constructing an intelligent tutor for the BUGGY game already exist in the form of the test validation techniques described in the previous section. Nevertheless, these techniques do not provide the right kind of information for explaining to a student *why* the problem he just generated had little or no diagnostic value. We are currently exploring the kind of reasoning required to answer "why."

An intelligent tutor designed specifically to help teachers will profit from a theory of what makes an underlying bug easy or difficult to diagnose. Simple conjectures concerning the depth of the bug from the surface do not seem to work, but more sophisticated measures might. It is hard to see how to predict the degree of difficulty in diagnosing a particular bug without a precise information-processing or cognitive theory of how people actually formulate conjectures about the underlying bug or cause of an error. For example, one such theory is that people walk through their own algorithm, looking for places where a part of the incorrect answer is different from their own and then try to imagine local modifications to their algorithm that could account for the error. Under this theory, one would expect bugs that involve major modification of the procedure, such as changing the direction in which columns are processed, to be difficult to diagnose. Similarly one would expect difficulty if a student's algorithm differs from the diagnostician's. Given an adequate theory, the difficult situations can be watched for and corrected through appropriate tutorial comments during diagnostic training.

---

[25] As a historical footnote, BUGGY was originally developed to explore the psychological validity of the procedural network model for complex procedural skills. During that investigation we realized the pedagogical potential of even this simple version of BUGGY as an instructional medium.

[26] For examples of the types of tutoring, see Brown and Burton (1975, 1977), Carr and Goldstein (1977), and Goldstein (1974).

Another extension to the BUGGY environment to encourage further exploration of the ideas of hypothesis formation and debugging is the development of a specialized programming language for writing simple arithmetic procedures. Actually having students write procedures provides immediate focus on debugging strategies—a topic usually left until the end in most secondary school programming courses. In this limited environment, it should be possible to construct an intelligent programming assistant as well as a computer-based tutorial helper that can aid a student when he gets stuck. Providing students with a language in which they can write their own procedures also allows use of a game developed in the SOPHIE environment (Brown, Rubinstein, & Burton, 1976), where one student writes a procedure introducing a bug and another student tries to discover it by presenting test problems.

Extensions to the Diagnostic System

Concerning the use of procedural networks as a tool for diagnosing real students, we reiterate that the capabilities of the present system are solely diagnostic: no tutoring is attempted. The issue of what tutorial strategy to use, even when it is known exactly what a student is doing wrong, is still an open question.

One possible strategy is that the "expert" portion of the procedural network could be made *articulate* in the sense of being able to explain and justify the subprocedures it uses. Since the system would know the student's bug, special problems can be chosen and special care can be taken while presenting the steps which illustrate the student's bug. This feature could also be used to allow students to pose their own problems to the system and obtain a running account of the relevant procedures as the "expert" solves the problem. A useful notion for the articulate expert may be to have additional explanation or justification of each symbolic procedure in the network expressed in terms of a "physical" procedure using manipulative tools such as Dienes' blocks (Dienes & Golding, 1970). In this way, the execution of each *symbolic* procedure could cause its analogous *physical* procedure whose execution could be displayed on a graphics device, thereby letting the student see the numeric or abstract computation unfold in conjunction with a physical model of the computation. This approach directly attacks the problem of getting procedures to take on "meaning" for a student: the acquisition of meaning is, we believe, accomplished by recognizing mappings or relations between the new procedures and existing procedures or experiences.

While we consider the articulation of the expert to have great promise as a corrective tutorial strategy, it is by no means the only possible such strategy. It is possible that with certain bugs (and certain students), a clear description of what is going wrong may be sufficient to allow the student to correct his problem. Or it may be possible to formulate a series of problems (possibly in conjunction with physical models) that enables the student to discover his own error. Or it may be best to abandon his algorithm (rather than trying to debug it) and start over with a different, simpler algorithm to build the student's confidence.

Generalization to Strategic Knowledge

Caution should be exercised in generalizing the procedural network model to other procedural skills. In particular, the aspects of knowledge discussed here are almost totally algorithmic in nature containing little heuristic or strategic knowledge in selecting or guiding the execution of the primitives (procedures). Many mathematical skills involve an interplay between strategic and algorithmic knowledge. For example, when adding two fractions, one does not necessarily compute a common multiple by multiplying together the two denominators. Instead, an examination is made of the relationships between the two denominators such as identity, obvious divisibility of one by the other, relative primeness, and so on. On the basis of such relationships specialized procedures are chosen for adding the given fractions. The rules underlying these decisions can have their own bugs and, therefore, these rules must be modeled within some representation scheme. Although procedural network schemes can represent such decision rules, we believe that other schemes, based perhaps more on annotated production rules (Goldstein & Grimson, 1977), deserve serious consideration. We are currently investigating a hybrid approach for modeling fractions (and their bugs) that involves merging annotated production rules with procedural networks.

Psychological Validity

An important area for exploration concerns the psychological validity of the skill decomposition and buggy variants in the network. That is, how well do the data structures and procedure calls in the network correspond to the structures and skills that we expect people to learn? From the network designer's point of view, the psychological validity can be improved or denigrated by choosing one structural decomposition instead of another. Determining a psychologically "correct" functional breakdown of a skill into its subskills is critical to the behavior of gaming and diagnostic systems using it. If the breakdown of the skill is not correct, common bugs may be difficult to model while those suggested by the model may be judged by people to be "unrealistic". For people playing BUGGY based on a nonvalid network, the relationship between its bugs and the behavior they observe in people will often be obscure. Measuring the "correctness" of a particular network is a problematic issue as there are no clear tests of validity. However, issues such as the ease or "naturalness" of inclusion of newly discovered bugs and the appearance of combinations of bugs within a breakdown can be investigated.

Finally, we have left open the entire issue of a semantic theory of how procedures are *understood* (and learned) by a person and why bugs arise in the first place. The need for a theory of how procedures are learned correctly or incorrectly is important for at least two reasons. First, an interesting theoretical framework that accounts for the entire collection of empirically arrived at bugs will undoubtedly provide insight into how to correct the teaching procedure that produced the bugs in the first place. Second, such a theory would be the next step in a semantically based generative theory of student modeling. As we have stated earlier, bugs have to be hand-coded into the network now. One can envision generatively producing bugs by a set of syntactic transformations (additions, deletions, transpositions, etc.) based on some appropriate representation language. While some bugs can be naturally accounted for as ''syntactic'' bugs, others, such as inappropriate analogy from other operations or incorrect generalization from examples, are best explained outside of the representation itself and hence require a ''semantic'' theory. One of the by-products of the diagnosis of the student data mentioned in the previous section has been a thorough and precise catalogue of bugs arising in one particular skill, subtraction. This network can now be used to suggest and evaluate theories about the origin of bugs.

APPENDIX 1: STUDENT RESPONSES TO BUGGY

In an experiment described in Section 2, a group of student teachers were exposed to the BUGGY game. This appendix provides a list of responses to the question, ''What do you think you learned from this experience?''

I see from this system that you learn from your mistakes. In a certain operation there are so many mistakes that you can make. When you learn what the mistakes are you learn to do the operation correctly.

That children's errors can be a way of diagnosing the way the child learns material. Also it raises questions about the way a child is tested both standardized and informally.

A student's errors and/or misunderstanding of a concept may have not been due to carelessness but rather involved a complex and logical thought process.

I learned that it is necessary to try many different types of examples to be sure that a child really understands. Different types of difficulties arise with different problems.

Although it's hard to tell from these pre- and posttests, in the middle is learned a great deal about the complexity of student's errors. I know that young students can get these preconceived notions about how to do things and it's very hard to find a pattern to their errors though it exists and I believe that BUGGY convinced me of [it].

That if you study the errors long enough you can *eventually* come up with a reasonable solution as to why the [error] is occurring.

Through looking carefully at children's math errors it is sometimes possible to discover a pattern to them. This pattern will tell you an area or a concept the child does not understand.

I learned that there could be more to a child's mistakes other than carelessness. Working with children with special needs I have encountered many such problems, yet never stopped to analyze what could be a systematic problem—for this I thank you.

Children do have problems and they are very difficult to spot especially when a number of different operations are used to come to an answer. I've learned to be more aware of how these children reach these "answers" and to help them to correct them; first by knowing how they arrived at the answer.

Although many arithmetic errors may be careless, there may also be a pattern that the kid is locked into. If you pick up on a pattern you can test the child to see if he/she conforms to it and work on it from there.

I found that I have looked closer at the problems, looking for a relationship between the set after working with BUGGY.

How to perceive problems, that don't look too consistent, a little easier.

I learned and was exposed to the many different types of problems children might have. I never realized the many different ways a child could devise to create his own system to do a problem. I am now aware of problems that could arise and I'm sure this will help me [in] my future career as a teacher.

How to more effectively detect "problems" students have with place value.

I have learned several new possible errors students may make in computation. I have also learned somewhat how to diagnose these errors, i.e., what to look for, and how specific errors can be.

I learned about diagnosing math difficulties. It makes me aware of problems that children have, and they sometimes think logically, not carelessly as sometimes teachers think they do.

That there are many problems that you can diagnose about a child by looking at his homework.

If a child has repeatedly made [the] same mistakes, it is more easily identified if the teacher has an opportunity to try and make [the] same mistakes. This method can be solved at least quicker than . . .

Tuned in to picking up malfunctions in simple addition and subtraction which seemed to be realistic problems.

ACKNOWLEDGMENTS

We are especially indebted to Kathy M. Larkin for her extensive assistance on this project and in particular for her contribution in refining the BUGGY activity and in discovering many new arithmetic "bugs."

REFERENCES

Barr, A. A rationale and description of the BASIC instructional program. Psychology and Education Series, Stanford University, Technical Report 228, April 1974.

Brown, J. S. Structural models of a student's knowledge and inferential processes. BBN Proposal No. P74-CSC-10, Bolt Beranek and Newman, Cambridge, Massachusetts, April 1974.

Brown, J. S., & Burton, R. R. Systematic understanding synthesis, analysis, and contingent knowledge in specialized understanding systems. In D. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science*. New York: Academic Press, 1975.

Brown, J. S., & Burton, R. R. A paradigmatic example of an artificially intelligent instructional system. Presented at the First International Conference on Applied General Systems Research: Recent Developments and Trends, Binghamton, New York, August 1977.

Brown, J. S., Collins, A., & Harris, G. Artificial intelligence and learning strategies. To appear in H. F. O'Neil (Ed.), *Learning strategies*. New York: Academic Press, 1978, in press.

Brown, J. S., Rubinstein, R., & Burton, R. R. Reactive learning environment for computer assisted electronics instruction. BBN Report No. 3314, A. I. Report No. 1, Bolt Beranek and Newman, Cambridge, Massachusetts, October 1976.

Brown, J. S., Burton, R. R., Hausmann, C., Goldstein, I., Huggins, B., & Miller, M. Aspects of a theory for automated student modelling. BBN Report No. 3549, ICAI Report No. 4. Bolt Beranek and Newman, Cambridge, Massachusetts, May 1977.

Burton, R. R., & Brown, J. S. Semantic grammar: A technique for constructing natural language interfaces to instructional systems. BBN Report No. 3587, ICAI Report No. 5, Bolt Beranek and Newman, Cambridge, Massachusetts, May 1977.

Burton, R. R., & Brown, J. S. A tutoring and student modeling paradigm for gaming environments. In *Proceedings of the Symposium on Computer Science and Education*, Anaheim, California, February 1976.

Carbonell, J., & Collins, A. Natural semantics in artificial intelligence. In *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Stanford University, 1973.

Carr, B., & Goldstein, I. Overlays: A theory of modeling for computer aided instruction. Massachusetts Institute of Technology, AI Memo 406, February 1977.

Collins, A., Warnock, E., & Passafiume, J. Analysis and synthesis of tutorial dialogues. In G. H. Bower (Ed.), *The psychology of learning and motivation*, Vol. 9. New York: Academic Press, 1975.

Cox, L. S. Systematic errors in the four vertical algorithms in normal and handicapped populations. *Journal for Research in Mathematics Education*, 1975, **6**, 202-220.

Dienes, Z. P., & Golding, E. W. *Learning logic, logical games*. New York: Herder & Herder, 1970.

Easley, J. A., Jr., & Zwoyer, R. E. Teaching by listening - toward a new day in math classes. *Contemporary Education*, 1975, **47**, 19-25.

Friend, J. Description of items to be used in addition/subtraction tests. Internal Memo, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1976.

Goldstein, I. Understanding simple picture programs. Massachusetts Institute of Technology, AI Laboratory, Technical Report 294, September 1974.

Goldstein, I. The computer as coach: An athletic paradigm for intellectual education. Massachusetts Institute of Technology, AI Memo 389, February 1977.

Goldstein, I., & Grimson, E. Annotated production systems. A model for skill acquisition. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, August 1977.

Krauss, R. M., & Glucksberg, S. Social and nonsocial speech. *Scientific American*, 1977, **236**, 100-105.

Matz, M. Some issues on building an Articulate Expert for high school algebra. BBN Report, ICAI 10 (in preparation), 1978.

Miller, M., & Goldstein, I. Overview of a linguistic theory of design. Massachusetts Institute of Technology, AI Memo 383, December 1976.

Rich, C., & Schrobe, H. E. Initial report on a LISP programmer's apprentice. Massachusetts Institute of Technology, AI-TR-354, December 1976.

Rubin, A. Hypothesis formation and evaluation in medical diagnosis. Massachusetts Institute of Technology, Artificial Intelligence Laboratory, AI-TR-316, January 1975.

Sacerdoti, E. *A structure for plans and behavior*. The Artificial Intelligence Series. New York: Elsevier North-Holland, 1977.

Searle, B., Friend, J., & Suppes, P. *The radio mathematics project: Nicaragua 1974–1975*. Institute for Mathematical Studies in the Social Sciences, Stanford University, 1976.

Self, J. A. Student models in computer-aided instruction. *International Journal of Man-Machine Studies*, 1974, **6**, 261–276.

Smith, M. J., & Sleeman, D. H. APRIL: A flexible production rule interpreter. SIGART Newsletter No. 63, June 1977.

West, T. Diagnosing pupil errors: looking for patterns. *The Arithmetic Teacher*, 1971, 64.

Young, R. M. Mixtures of strategies in structurally adaptive production systems: Examples from seriation and subtraction. SIGART Newsletter No. 63, June 1977.

TOWARDS AN EFFECTIVENESS MEASURE IN READING

Bertram L. Koslin, Sandra Koslin and Susan Zeno
Touchstone Applied Science Associates (TASA)*

I. Introduction

A. The Need for a Measure of Effectiveness in Reading

Current measures of reading achievement do not directly assess educational outcomes in that they do not measure how well or how much students can actually read of the corpus of available materials in English. It may be argued that conventional tests do measure outcomes, inasmuch as it is possible (especially in areas other than English comprehension) to show that following instruction (e.g., in French) students are able to correctly answer questions which they were unable to answer prior to instruction. However, the assignment of meaning to given scores (either the number right or percentile scores) on tests of different skills is left as an exercise for the consumer. What, for example, is the functional meaning of having passed a criterion referenced test? Does either of these scores mean that a person can read a newspaper in French? Perhaps, as Angoff (1971) notes, those people who process such scores over a period of many years do develop a functional meaning for them. However, most consumers of test scores should not be expected or required to develop such a base of experiences.

Mastery tests measure the attainment of discrete instructional objectives. Norm-referenced test scores reflect the comparative standing of students in a standardization group. But none of these scores provides the information needed to conclude that some verified progress has been made towards an "end objective." With present measures, for example, we are unable to directly ascertain whether or not schools are preparing students to function competently in meeting the reading requirements of the adult world.

The type of reading test that is required to measure adequately what has been learned as a result of receiving educational treatments is one that differs in a number of significant respects from any measure heretofore available. Therefore, it is useful to assign a generic name to the new type of measure. We have selected the name "effectiveness measure" to refer to a measure of outcome which provides a directly interpretable assessment of the extent to which an enduring adult ability such as reading with comprehension has been attained.

An effectiveness measure in reading must have certain properties. It must provide functional information concerning reading ability. In other words, it must be a test whose scores are directly interpretable with respect to how well, or what, students are able to read. Since we assume that the ultimate objective of all reading instruction is to prepare students to

---

*The research reported in this paper has been sponsored by The Board of Regents of the State of New York and the Carnegie Corporation of New York.

function competently at the reading tasks of the adult world, the measure should be able to provide information concerning whether progress is made toward that objective.

With long-term measurement of growth toward adult reading competence in mind, the effectiveness test should be applicable over a wide age range. It must be sensitive to group growth in achievement and, preferably, to individual growth in reading achievement as well.

B. Norm-Referenced Tests: An Unsatisfactory Approach to the Measurement of Effectiveness

The fundamental barrier to using standardized norm-referenced tests as effectiveness measures in reading is that they do not directly measure what students have learned. Attempts have been made to give rough functional meaning to norm-referenced scores, e.g., to use the eighth grade norm as a loose definition of competence. However, in order to assign a precise functional meaning to a test score, the score must have only one interpretation. Scores on norm-referenced tests (as well as on other types of tests) can be earned in many ways, since they are based simply on the number of heterogeneous test items answered correctly, without regard for which items were answered correctly. For example, on reading tests, there is simply no theory (see Section C) with which to interpret the functional significance of obtaining various combinations of correct answers on items designed to measure vocabulary knowledge, understanding of main ideas, literal and inferential comprehension, etc. A score on a norm-referenced test therefore has too many different meanings to be given a clear functional interpretation in terms of what students earning that score are able to do.

Another major problem in using norm-referenced reading achievement tests as outcome measures is that it is often not clear what is being measured. Careful examination of the items on such tests raises many questions concerning which skills are actually assessed. Since test publishers do not provide any operational definition of what reading comprehension means,[1] it is hard to evaluate whether or not a test measures what it is supposed to measure. The fact that it is possible to answer -- at significantly better than chance levels -- the questions on major conventional reading achievement tests without reading the accompanying text (e.g., see Tuinman, 1974) must raise questions concerning which student accomplishments the tests are actually measuring.

Furthermore, while it is true that technically adequate standardized tests measure achievement more of less reliably by discriminating among students' ability to answer a particular set of questions at a given point in time, standardized tests are not sufficiently reliable to measure growth -- the main concern of education. Because of the highly unreliable nature of score differences, individual gain scores are nearly meaningless.[2] Thus norm-referenced tests do not meet another of the criteria for an effectiveness measure, namely, that it be capable of measuring student growth over time.

Thus, for several reasons, norm-referenced tests represent an unsatisfactory approach to the measurement of effectiveness. Clearly they are unsatisfactory as effectiveness measures because they do not directly measure what students have learned in terms of some sought-after (socially validated and valued) outcome. Since norm-referenced achievement tests are designed to discriminate among individuals, they cannot be expected to be sensitive measures of educational outcomes. In addition, it is not always clear what norm-referenced tests are actually measuring. Certainly no definition of what is meant by reading with comprehension is stated in conjunction with operational procedures for measurement.

C. Criterion-Referenced Tests: Measures of Outcome but Not Effectiveness

Standardized norm-referenced tests of achievement dominate the reading measurement field, but certainly they are not the only tests available. Most notably, there has been a recent movement in testing toward the development of mastery or criterion-referenced tests. Such tests are designed "to yield measurements that are directly interpretable in terms of specified performance standards" (Glaser & Nitko, 1971, p. 673). In clear distinction to norm-referenced tests, mastery tests are direct outcome measures. However, it would be difficult to employ the usual criterion-referenced tests as effectiveness measures.

Criterion-referenced tests are typically (although they do not necessarily have to be) limited in scope to the mastery of specific units of instruction. While the tests may adequately measure whether students have or have not mastered some particular instructional objective (such as learning to blend initial dipthongs), they are not good measures of complex long-term educational outcomes, such as having learned to read well enough to function as a competent adult.

The fact that criterion-referenced tests typically deal with specific units of instruction does not by itself invalidate them as an approach to measuring effectiveness. The problem arises from the fact that there is no known empirically verifiable way to aggregate results across criterion-referenced tests to measure growth toward adult competence. If a criterion-referenced test approach were to be used to measure a person's ability to comprehend materials at a level that characterizes adult competence, it would be necessary to know how to weight and aggregate performance over various validated skills which demonstrably comprise reading comprehension.

At present, lacking a validated model of reading comprehension, we neither know the component "skills," nor how to assign the weights, nor how (from an empirical point of view) to aggregate performance over the "skills." Consequently, we are constrained, at present, in trying to use the results of several criterion tests to arrive at a single judgment of competence.

It follows from the above discussion that criterion tests are presently restricted to two methods for reporting results. The first method is simply to describe and display the test results without drawing any conclusions concerning the attainment of competence. This method of presentation is seen, for example, in reports of the National Assessment of Educational Progress. Performance of students on individual exercises (items) is presented in terms of the proportion of students nationally and in each group (defined by race, sex, geographic region, etc.) who get the exercise correct, and in terms of group deviations from the national proportion of success. Data are summarized by pooling related sets of exercises to determine the median deviation from national performance of each group, and then comparing overall performance in terms of the deviations of the various groups. No total (or average) scores are reported across exercises, and no interpretations are made regarding competence or attainment of standards.

The second method of reporting results from criterion-referenced tests goes one step further by having judges set criteria of competence. An illustration is provided by the Adult Performance Level Study (1973). Instead of simply displaying the proportion of respondents answering each question (performing each task) correctly, performance on related tasks was aggregated and expert judgment used to classify performance as functionally illiterate, functional but not proficient, and proficient.

In order to produce a genuine effectiveness measure using the typical kinds of criterion-referenced tests, a well-validated model of reading comprehension, which is lacking at present, would need to be developed. Only after such a model was developed would it be possible to take a meaningful weighted average of particular reading skills as measured by the various tests. To develop such a model it would be necessary to enumerate all reading skills, validate them (show that they were instrumental to reading with comprehension), and establish the empirical relationships of each skill, individually and in all possible interactions, to the outcome behavior of reading comprehension.[3] Necessarily, a rigorous operational and independently validated definition of the construct "reading comprehension" would have to be provided in order to weight the component skills.

Studies would be required not only to determine the contribution of each skill to reading comprehension, but also to learn which skills impact directly on comprehension and which have indirect facilitating effects (serve as moderator variables). Since the role of each subskill as a direct or a moderator variable in reading comprehension (as well as its relative importance) may change as a function of age or reading development, it would be necessary to study the relationships between subskills and comprehension developmentally as well as for adult readers.

In short, building a model of reading comprehension -- a prerequisite to using conventional criterion-referenced tests as effectiveness measures -- would require work in at least four areas: (1) identifying and validating all the reading subskills; (2) operationally defining what is meant by reading comprehension; (3) showing that reading with comprehension does not occur in the absence of a subskill and (4) establishing the relationships between subskills and comprehension so that the proper weights and aggregation rules could be assigned (at different stages of reading development).

The approach that we are proposing for formulating effectiveness measures aims instead at measuring the outcome performance (i.e., reading comprehension) directly, without weighting and aggregating component skills. This implies the absolute necessity to define what is operationally meant by reading comprehension, but eliminates the need to identify all reading subskills as well as the need to investigate relationships between the subskills and reading comprehension. Hence the proposed approach permits us to arrive at an effectiveness measure more directly and more rapidly than would a conventional criterion-referenced test approach. (We note, in passing, that once the effectiveness measure was developed, the measure itself could be used in the construction of a model of reading comprehension and its acquisition.)

D. Design Concepts for a Distribution-Free Effectiveness Measure

The approach that we are proposing for the development of an effectiveness measure in reading is new to educational measurement. As the explication below will reveal, the approach is clearly not norm-referenced, and while it shares certain properties with objective-referenced tests, it differs from them in critical respects.

1. Desired properties of the effectiveness measure. Our approach to the measurement of effectiveness calls for the development of a test (or family of tests) with the following distinguishing properties.

1.1 The test should be based on an operational definition of reading comprehension. The definition of reading comprehension that is used in building the test should be empirically validated to show that it is a meaningful construct, circumscribing the phenomena to which we refer when discussing "reading with comprehension." In addition, the definition should have utility in the formulation of theory, for example in building a model of how reading comprehension ability is acquired.

1.2 The test should measure different levels of comprehension on a common scale that has at least interval and, preferably, ratio properties.

1.3 The test should be sensitive to growth in reading achievement over academically relevant periods of time, such as one year. Sensitivity to group growth is a requirement, and sensitivity to individual growth is highly desirable.

1.4 The test should have scores that can be functionally interpreted with respect to what students are able to do, at least on ecologically important reading tasks in "real life;" therefore, there should be as few ways as possible -- preferably only one -- to earn a given score.

2. Reading comprehension and domain-referenced test items. In response to the need for reading tests that measure what students know or are able to do, many competency and performance tests have appeared in the field. However, the current tests focus too much attention on the mode of referencing scores and too little on the test items themselves. The current tests are made up of items not unlike those that have appeared in reading tests for many years. Criterion-referencing (or norm-referencing) these typical "types" of multiple-choice test items does not provide any information concerning how well a person performs the cognitive process that we call reading with comprehension.[4] In order to measure how well a student reads, test items are needed which correspond to a process definition of reading with comprehension.

Such a definition has emerged from recent work in psycholinguistics. From this point of view, reading is a complex act, the main purpose of which is extraction of meaning. The process is rooted in the brain's capacity to acquire and organize incoming information (in the case of reading, the ability to uptake and organize the visual configurations in text). According to psycholinguists, the reading process is active and it is selective. The fluent reader utilizes syntactic and semantic cues from the text, along with prior knowledge of language and content, to reduce uncertainties concerning meaning and to confirm or disconfirm predictions. Comprehension can be defined in terms of the ability to make use of (process) information provided in text to decide which of several alternatives best conforms to the (socially accepted) meaning of the text. A student can be considered to "comprehend" when (s)he has processed the available information to the point of having eliminated all but the "correct" meaning of the text.

The psycholinguistic point of view leads logically to specifications for an item to measure comprehension. The essential characteristic of such an item is that answering correctly be contingent upon processing prose. One way to achieve such an item is to delete selected words from text and provide response options with which the student is to restore the text such that processing surrounding prose is both necessary and sufficient to choosing a right answer.

3. The organizing concept: the difficulty (readability) of prose. Adopting a psycholinguist definition of comprehension at the level of items is not by itself sufficient to developing an outcome measure of achievement in reading. The definition leads to a set of items, but provides no way of organizing them to obtain a measure of ability or to measure growth toward an end objective. For this purpose a second construct, namely a scale of measurement, is required. For this we propose to use readability or, more precisely, the difficulty of prose.[5]

Use of a difficulty of prose scale to organize a test begins with a set of verifiable assumptions. First, we assume that the difficulty of prose varies. Further, we assume that variation in the difficulty of prose is a function of processing requirements: as the processing demands of prose increase, the prose grows more difficult. Finally, we assume that processing requirements are accounted for by certain structural features, and that they can be indexed or quantified accordingly.

If we can develop a test item that is linked to the information processing requirements of material, then item difficulty should vary with the difficulty of the prose, i.e., as information processing demands increase, items should get harder. Therefore, we should be able to build a measure of the ability to process harder and harder prose if we array prose in ascending order of difficulty, holding constant what the items require subjects to do. Our ability to build such a test is contingent, first and foremost, on developing an item that is so dependent on the processing requirements of prose text that its difficulty will be directly proportional to the difficulty of that text.

The use of prose difficulty as an organizing concept has important consequences. It makes it possible, on the one hand, to define reading tasks quantitatively and, on the other hand -- if prose difficulty is made an element of test design -- to define ability in terms of the difficulty of materials that can be read. Therefore we can quantitatively relate performance to task demands.

4. A statement of propositions. The propositions underlying the work may be stated as follows:

a. If the difficulty of prose can be scaled, and

b. If an item can be built such that responding correctly at any level of prose difficulty depends entirely on processing surrounding prose at that level (i.e., if processing text is a necessary and sufficient condition for responding at any level of difficulty), and

c. If ability is defined in terms of the difficulty of the items passed, such that the harder the items passed the higher the ability,

Then

d. A measure of the difficulty of prose can be translated into a requirement of ability for processing that text (more simply: knowing how hard the text is, we can specify the ability that someone will need in order to process it).

e. Corollary: A measurement of ability can be translated into a forecast of the difficulty of text that can be processed (more simply: if we know a person's ability score, we can predict how likely (s)he is to understand various ordinary prose selections).

5. Satisfying the propositions. Each of the three propositions underlying the work can be and/or has been satisifed.

5.1 The difficulty of prose can be scaled. Variation in the difficulty of prose is a property of languages whose existence and importance is acknowledged by researchers and practitioners from many schools of thought. Behaviorists (e.g., Skinner, 1957), psychometricians (e.g., Davis, 1968), psycholinguists (e.g., Smith, 1973), reading specialists (e.g., Chall, 1967) and teachers (e.g., Cooper & Petrosky, 1976) all recognize that written (as well as oral) materials vary in difficulty.

5.1.1 Empirical verification that prose materials form a difficulty scale. Since the systematic study of readability began early in this century, various methods have been employed to measure the difficulty of prose passages. This research has provided empirical verification, across a variety of methods, that prose materials do in fact form a difficulty scale. Textual difficulty has been evaluated by direct judgments of students (e.g., Washburne & Vogel, 1926), teachers, and various experts (Hayes, 1934; Smith, 1930). The method of paired comparisons, a related procedure where every passage to be judged is compared with every other, has also been used (Hackman & Kershner, 1951). Still other investigators (e.g., Kyte, 1925; Mathews, 1926; Staiger, 1955) have employed direct tests of comprehension, speed, error rates or similar indices to measure the difficulty of text. Regardless of method, investigators have found that they could array prose passages from easy to difficult.

Passages have likewise been arrayed for difficulty in terms of student responses to multiple-choice questions asked about them (e.g., McCall & Crabbs, 1925) and in terms of cloze test scores for the passages (Bormuth, 1966; Coleman, 1965). Over the years it has been found that when passages are so arrayed their difficulty can be statistically accounted for by a number of quantifiable features of the language. These relationships between structural features and difficulty have been used to develop "readability formulas" whereby the difficulty of new material can be predicted by analysis of the same set of features.

The different formulas so developed tend to agree in rank ordering the difficulty of "new" prose passages.[6] Furthermore, the ratings of text difficulty by the formulas are similar to those resulting from other methods, such as direct judgments (see Klare, 1963 for a review of this literature) or cloze techniques (Taylor, 1953). Essentially similar predictions are made concerning the relative difficulty of passages of text, whether difficulty is measured by formulas based on traditional counts of mechanical features or on relatively sophisticated contemporary linguistic analysis (e.g., Mac Ginitie & Tretiak, 1971). Thus in the aggregate there is compelling evidence that passages of text can be arrayed in terms of difficulty. We may accordingly assume that prose materials do form a readability scale.

5.1.2 Verification that structural variables account for readability. Fifty years of research have identified certain structural variables that account for the readability of prose. Factor analysis of those characteristics that have been the best predictors of readability has identified two major factors: vocabulary difficulty and sentence complexity (Brinton & Danielson, 1958; Stolurow & Newman, 1959). Of the two factors, vocabulary difficulty has consistently been the more important predictor. Difficulty of vocabulary, defined by the presence or absence of words appearing on a list of "easy" (familiar) words, has been shown to account for anywhere from 26% to 46% of the variance in the readability of a set of criterion passages (Lorge, 1948; Dale & Chall, 1948; Spache, 1953; MacGinitie & Tretiak, 1971; Bormuth, 1966). In one study (Coleman, 1971) vocabulary difficulty alone accounted for 83% of the variance in a set of criterion scores of readability.

Word length (in letters or syllables), another index of vocabulary difficulty, has been found to account in recent studies for some 46% to 81% of the variance in passage readability (Bormuth, 1966; Coleman, 1971).

Sentence length in words, the easiest to compute and most frequently used index of sentence complexity, correlates around $\underline{r} = .5$ with various criteria of readability (Lorge, 1948; Dale & Chall, 1948; Flesch, 1948; Coleman, 1971; Bormuth, 1966). Sentence length in letters and syllables may have even higher correlations (around $\underline{r} = .7$) with passage readability (Bormuth, 1966).

Other structural variables that have been identified as accounting for readability include clause length, number of prepositions, and prepositional phrases, and mean word depth (Gray & Leary, 1935; Lorge, 1948; Bormuth, 1966).

5.1.3 Analytic verification of structural variables. Many of the structural features of language that have proven useful as predictors of textual difficulty can themselves be arrayed on a scale of difficulty, familiarity, complexity, or another similar dimension. Vocabulary (or word frequency) is log-normally distributed (Herdan, 1960; Carroll, 1968). A very few words in the lexicon have high frequencies of occurrence, but most words appear infrequently. Research supports the intuitively reasonable assumption that those words occurring frequently in written language will be more familiar to readers than will infrequently appearing words. It has been found, for example, that reading rates are faster for more frequent than for less frequent words (Pierce & Karlin, 1957); that words with high frequencies of occurrence are recognized more rapidly than words with low frequency of occurrence (Howes & Solomon, 1951) and that they are heard more readily in noise (Postman & Rosenzweig, 1957). Thus there is empirical evidence that the mathematical scale on which vocabulary can be arrayed has psycholinguistic significance to users of the language.

Word length, a major predictor in some recent formulas, tends to bear an inverse statistical relationship to the relative frequency of words (Zipf, 1935). And sentence length, another important predictor variable, has been found to be highly correlated with several indices of syntactic complexity, including rather sophisticated ones based on recent developments in psycholinguistics (e.g., see MacGinitie & Tretiak, 1971). Thus a difficulty (or familiarity, complexity, or similar dimension appears to be characteristic of many subsidiary features in the language as well as of the composite language itself.

In sum, there is convergent evidence that prose difficulty can be scaled, that structural variables account for it, and that difficulty is a psycholinguistically valid way to quantify prose reading tasks.

5.1.4 *Ordering the difficulty of prose passages*. Readability formulas index prose difficulty. Therefore, one or more of the available formulas can be used *a priori* to organize a power test. Prose passages can be sequenced so that progressively more difficult text is encountered from the beginning to the end of the test. Later (i.e., after the test has been administered) the ordering on the test can be analytically verified.

5.2 *An item-type to satisfy the second proposition*. It is possible to satisfy the second proposition: namely, that an item be built for which responding correctly at any level of prose difficulty depends entirely on processing surrounding prose at that level. The proposition is satisfied by a particular item type that operationalizes a process definition of reading comprehension by deleting selected words from text and providing response options with which the student is to restore the text, such that processing the surrounding text is necessary and sufficient in order to respond correctly. This item type has the following characteristics:

5.2.1 *The item is text dependent*. The item cannot be answered unless the surrounding prose has been processed and understood.

5.2.2 *Answers to items depend on processing the intersentential context*. Information must be integrated across sentences in order to answer the item correctly. The syntactic and semantic information required to select the correct answer to an item is located in the intersentential context surrounding the sentence frame containing the test item, and *not* in the test item frame itself. There is insufficient syntactic and/or semantic information within the sentence frame of the test item and/or the response options to allow the student to select a correct response. This is a necessary condition for interpreting test performance as meaning that a student has read and understood continuous prose text.

5.2.3 *Ability to understand the prose is not confounded with ability to understand the responses*. The difficulty of test items should be linked to the difficulty of the intersentential context, not to the difficulty of the deleted words. Accordingly, the familiarity of those vocabulary words which may be deleted (or used as wrong response options) is controlled at the level of highly familiar (frequent) English words. A test composed of such items get harder only with respect to the continuous text students must process. Thus, correctly answering test questions measures comprehension of continuous prose text at increasing levels of difficulty , and this ability measure is not confounded with the vocabulary difficulty of the response options.

5.2.4 *The prose is informationally adequate*. Because comprehending the test must be a *sufficient* condition for answering questions correctly, all the information needed in order to choose a correct response is provided internally within the passage. Therefore, no prior experience with subject matter content is required.

5.2.5 *The correct answer is obvious if the prose is understood*. Response options are provided such that the right answer is obvious if the prose has been processed and understood. The right answer is unambiguously correct, and all others are unambiguously wrong. The item if free of extraneous cues (unrelated to comprehension of the continuous prose) on which students might base guessing strategies.

5.2.6 Items are independent. Every item is linguistically independent. Responding correctly is not contingent upon having answered another item correctly, but only on processing the prose context surrounding the item itself.

5.3 Satisfying the third proposition: that ability be defined in terms of the difficulty of items passed. If a true power test is organized (i.e., if the items get harder from the beginning to the end of the test and do so because they demand more of the ability in question, rather than for irrelevant reasons), then the harder an item is, the higher must be the ability of the student passing it. Thus ability can be defined in terms of item difficulty, such that the harder the items passed the higher the ability.

6. An operational definition of reading comprehension. The task particulars specified above result in an operational definition of reading comprehension. Reading comprehension is defined as the ability to use the syntactic and semantic information provided by the intersentential context in prose at different levels of difficulty to correctly complete (restore) a missing section of the message, using a set of response options controlled for word familiarity, under conditions where the text is informationally adequate to answering the question and there is only one correct answer. The student's ability score is defined as the probability of correctly answering items of various difficulties which are linked to levels of readability.

Relative to the psycholinguistic definition of reading comprehension, this operational definition, embodied in the present Degrees of Reading Power Test, is only one of a class of possible operational definitions. Nevertheless, it results in a test where the conditions of measurement are operationally defined, a critical step in building an outcome measure.

E. Criterion-Referencing the Test

If the three conditions of measurement are satisfied (see D4), then knowing ability, we can forecast the difficulty of prose that can be understood. Conversely, knowing the difficulty of prose text, we can predict the ability required to process it. This means that it is possible for an effectiveness measure in reading to be objectively criterion-referenced in terms of predicting performance on tasks external to itself.

The heart of our approach to the development of an effectiveness measure in reading is to use the scaled difficulty (readability) of text -- an important property of language -- both as the basis for defining standards of reading competence and as the basis for measuring progress toward the attainment of those standards. If the written materials that students or graduates of the educational system are expected to read are specified, and if these written materials are scaled for readability and the values appropriately aggregate, then the level of reading comprehension ability that students or graduates of the educational system must attain is operationally defined by the "average" measured readability characteristics of the designated materials. Thus, performance standards for educational processes can be established.

Once various quantitative standards of reading competence are defined empirically, simple and direct measurements of the extent to which students have attained these standards can be made by administering reading tests designed to determine whether individuals can read with comprehension passages that have been scaled for the same linguistic properties that were used to define the standards, i.e., for readability.

Our approach to the construction of an effectiveness measure in reading follows logically from the preceding argument that standards of reading competence can be defined in terms of the scaled, linguistic properties of reading materials. It calls for a test in which passages vary in readability (e.g., from easy to difficult). A student's performance on passages whose readability corresponded to that of a performance (e.g., adult) standard would be directly interpretable in terms of whether of not that standard of competence had been met.

Because all test passages would be described quantitatively on a common scale of readability, performance on one test could be related directly to performance on another test. It follows that if the test passages were made to reflect the full range of school reading materials -- from primers all the way to adult-level text -- then periodic administration of the test could be used to monitor a student's progress toward adult competence as (s)he moved through school.

The materials that are assigned to students to be mastered during a school year constitute a set of expectations or de facto standards for reading achievement. By analyzing the assigned instructional materials at each grade level to determine their readability, it is possible to define these de facto standards quantitatively. The effectiveness measure can then be used to measure attainment of these de facto grade level standards: it can be used to compare the difficulty of text that students are able to understand with the difficulty of text they are expected to understand, and hence can indicate the extent to which each student meets grade level expectations at the beginning and end of each school year.

In addition to predicting whether or not some fixed (e.g., adult or grade level) standard has been met, the effectiveness measure can be criterion-referenced with respect to performance on any prose task that can be scaled for readability. Knowing a person's ability, and knowing the difficulty of the task, a prediction of that person's probable success on that task can be made. Being able to criterion-reference the test in this way has implications for the management of instruction. If we accept the traditional wisdom that instructional materials should be suitably difficult relative to student ability (e.g., that 75% comprehension defines the "instructional" level and 90% the "independent" level), then ability scores on the effectiveness measure can be used to predict which prose materials will be suitable for different students for purposes of instruction, and which will be suitable for independent work (assuming that the materials have been scaled for difficulty).

From the point of view of managing instruction, the setting of standards or expectations for individuals also involves knowing the optimal discrepancy between a student's current level of ability and the difficulty of instructional materials which will maximize growth and motivation while minimizing the development of negative affect such as frustration or self-deprecation. Bormuth's (1971) work on defining such optimal discrepancies between ability and instructional materials, while not conclusive, illustrates that such expectations (one for each level of ability) can be developed.

II. The Formulation and Validation of an Effectiveness Measure in Reading

Considerable progress has been made toward the formulation and validation of a test of prose comprehension with the properties specified in Part I. The test, which is called the Degrees of Reading Power (DRP) test, fulfills the requirements of an effectiveness measure as previously outlined. Its distinguishing characteristics can be summarized briefly, as follows.

The DRP is a test of prose reading comprehension designed to determine the hardest non-technical expository prose that a student can understand on the "naturally" continuous scale of prose difficulty (readability). In accord with a psycholinguistic view of the reading process, the DRP defines reading comprehension as the ability to use syntactic and semantic information provided by the intersentential context in prose to derive sufficient meaning(s) to correctly complete a missing section of a message. The DRP is designed to measure this ability on any non-technical expository test, at different levels of prose difficulty.

Every item in the test embodies this definition of comprehension. Each DRP test item is formed by deleting a content word from text and asking subjects to restore the deleted word correctly by selecting the appropriate response from among several alternatives provided. Rules of formation for the test items are such that processing the text in surrounding sentences in the passage is necessary and sufficient for responding.

Processing the prose in the passage is necessary for responding correctly because all the response options for a given item fit the blank space: each makes a correct and plausible sentence when inserted in the blank if the sentence is considered in isolation. However, only one response is plausible when the surrounding context of the passage is considered. Therefore, to select the correct response, the student must read and understand the prose in the surrounding sentences of text. If (s)he does, then the one correct answer will be obvious.

Processing the prose is sufficient to responding correctly because the test is informationally adequate: Enough information is provided within the passage to allow a student to choose a correct response. No prior experience with subject matter content is required. Therefore, a student who has read and understood the text should be able to answer the questions.

To ensure that it is the ability to comprehend continuous prose at given levels of difficulty that is being measured, the difficulty of the deleted words and the response options is held constant at the level of highly common words throughout the test. The ability to respond at any level of readability therefore depends on the ability to process prose at that level, and is not confounded with the difficulty of the particular word deleted or the difficulty of the response options.

The DRP is organized as a power test. It is composed of a series of expository prose passages, each of which contains seven items with the characteristics enumerated above. The passages are arranged in ascending order of difficulty. The Bormuth mean cloze formula (Bormuth, 1969a) is used *a priori* to sequence the passages so that progressively more difficult prose is encountered from the beginning to the end of the test. After the test has been administered, the order of the readability of the passages is analytically verified. Present DRP tests are broad-band, measuring the ability to read text ranging in difficulty from first to graduate level prose.

The DRP is scored to provide a sample-free (i.e., non-normative) estimate of students' reading ability. Conceptually, the DRP may be thought of as comprising a series of hurdles of increasing difficulty (the ordered series of passages). According to the theory of the test, the student correctly answers items (clears the hurdles) until (s)he comes to a point where (s)he can no longer understand the passage sufficiently well to grasp the context surrounding test items, and hence can no longer decide which option is correct. At this point (s)he has reached the limits of his (her) reading ability and will no longer clear the hurdles.

A scoring model which corresponds to the concept of the DRP is the Rasch model, a latent trait model with two parameters: the ability of the person taking the test and the difficulty of the test items. According to the Rasch model, the probability of a correct response to a test item is a geometric (logistic) function of the difference between the difficulty of that item and the ability of the person responding. The model provides the mathematical bases for calibrating both items and people in terms of difficulty and ability, i.e., for obtaining unbiased estimates of the model parameters, and for determining the probable success on items for persons of different abilities. The logic of the Rasch model is consistent with the design and structure of the DRP. That is, individuals may be expected to perform successfully on all items (or passages) well within their ability, perhaps show inconsistencies in performance as they approach the limit of their ability, and fail all items (or passages) having difficulties well beyond their ability.

Since other investigators have sometimes discarded items which did not fit the Rasch model (Wright, 1977), it is important to emphasize that in the case of the DRP the model was adopted because it fit DRP data *without* any items needing to be altered or discarded. Explicitly, the DRP was not built to fit the Rasch model. The use of the Rasch model does not "add" or "subtract" anything concerning measurement with the DRP: the requirements for formulating DRP passages and test items are not affected by use of the model.

A variety of studies and experiments were initiated to evaluate whether the DRP is a valid and reliable test. For these studies and experiments, twenty-four passages were prepared, each containing seven items. These 24 passages comprised two passages at each of 12 levels of difficulty ranging from extremely easy prose (about first grade) to quite difficult prose (about college-level). The two passages at each level of difficulty were then randomly assigned to one of two test forms. Approximately 5,000 fourth and sixth grade students, drawn from schools in New York State where the minority population varied from 20 to 80 percent, were assigned at random to participate in one of several studies, which are briefly reported below. These studies are reported in greater detail elsewhere (see Koslin, Koslin, Zeno, and Wainer, 1977).

A. Traditional Reliability and Validity Studies

Students took the two randomly constituted "parallel" forms of the DRP (see preceding paragraph) with at least a week's interval between the administration of the forms. The order of administration was counter balanced. Product-moment correlations showed that the DRP is reliable according to the constructs of True Test Score theory. For the randomly constituted "parallel" form, $\underline{r}$ = .90 using Rasch ability scores. The DRP is also homogeneous, the a KR_{20} = .95 for both forms. Thus by conventional standards, the DRP is reliable.

The DRP also meets conventional standards for validity inasmuch as both "convergent" and "discriminant" evidence were found for the construct validity of the test. In different studies, students took either the DRP and a norm-referenced reading achievement test, or they took the DRP, two kinds of cloze tests, and an IQ test. Results of these studies have showed that the DRP correlates more with the language subtest of an aptitude measure ($\underline{r}$ = .80) than it does with the non-language subtest ($\underline{r}$ = .70). The DRP, as expected, has moderately high correlations with an "any-word" cloze test ($\underline{r}$ = .76), with a "hard" cloze test ($\underline{r}$ = .78), and with a conventional norm-referenced reading test ($\underline{r}$ = .84). In accord with expectations, the DRP is less strongly related to a measure of vocabulary knowledge of words presented in partial isolation ($\underline{r}$ = .56) than it is to traditional multiple choice comprehension questions ($\underline{r}$ = .80).

B. Nontraditional Validity and Reliability Studies

A set of studies were designed to evaluate whether the DRP is valid and reliable in accord with the meanings of these terms in the natural sciences.

1. *Construct validity and the empirical meaning of DRP passage difficulty scores*. A number of studies were carried out to determine whether the average item difficulty per DRP passage is accounted for by the difficulty of the surrounding prose.

In one experiment the issue was whether the difficulty of DRP items would change when the difficulty (readability) of the surrounding text was altered. The fact that the difficulties of identical items (test sentences and reponse options) varied as a function of the difficulty of the prose in which they were embedded suggested the DRP item difficulties measured the readability of prose.

In another study, an attempt was made to determine whether surface linguistic variables known to account for readability could predict DRP passage difficulties. The mean and median Rasch item difficulties were calculated for each passage. These average passage difficulties became the criterion scores in two multiple correlations. The predictor variables were the proportion of hard words, the average word length, and the average sentence length per passage. The multiple R's between the predictor variables and mean and median passage difficulty were, respectively, $R = .98$ and $R = .96$ for one form, and $R = .95$ and $R = .96$ for another form. Thus, essentially all the variance in average passage difficulty was accounted for by the surface variables. As best as can be measured, the mean or median Rasch item difficulties have no meaning other than indexing the difficulty of prose.

Finally, a correlational study showed that the mean and median DRP passage difficulties were essentailly accounted for by Bormuth's (1969a) mean cloze forecasts. The more or less sample-free DRP passage difficulties for both forms correlated ($r = -.98$) with Bormuth's (1969a) measure of readability.[8]

The near perfect correlations between the average item difficulties of the DRP passages and the Bormuth readability values for the passages mean that the DRP can validly be criterion-referenced, and that it is also possible to make valid predictions concerning performance on prose reading tasks external to the test itself. One regression equation allows the DRP to be criterion-referenced if standards are formulated in terms of readability requirements. Once such a standard is formulated, then the ability required to comprehend at that readability level can be predicted. The other equation enables the DRP to forecast the likelihood that students of different abilities will be able to read prose at given levels of readability.

2. *Predictive validity: Forecasting the likelihood that subjects can comprehend prose at given levels of difficulty*. Manipulations with the Rasch model enable the probabilities of success on prose reading tasks for students of different (DRP) abilities to be predicted. The validity of these procedures depends upon the fact that the Rasch model fits DRP data, and the fact that average DRP passage difficulties measure readability. Even so, a major uncertainty had to be reduced: do DRP passage difficulties measure readability only for the particular (e.g., informationally

complete) expository prose of the DRP or do they forecast what students can do with other prose tasks? This uncertainty was reduced in an experiment, as follows.

The DRP was administered to students, and ability groups were constituted by predicting which subjects would be at the instructional level (i.e., have a .75 likelihood of correctly answering DRP-type questions) on each of six prose passages in a set arranged for increasing difficulty. When forecasts were compared with the actual success of students, the DRP did predict the traditonal (p = .75) instructional level. It also predicted which students would have a .90 likelihood of success (commonly referred to as the independent level) with materials of different levels of readability. Other predictions were also confirmed. Thus it seems that the DRP can predict how much information will be acquired or gained from ordinary prose.

Parenthetically, one prediction that was confirmed is particularly important in using the DRP to match materials with student ability. We were able to show that giving students intact text to read just before they took a comprehension (hard cloze) test led to "mastery" level performance if and only if the readability of the text was in-range relative to subjects' ability as predicted by the DRP (i.e., if the probability of success equalled p = .75). Thus, if the aim of instruction is to enable individuals to master materials in terms of assimilating their content, it would appear that the conventional wisdom elaborated several decades ago is correct: no more difficult materials than those falling within an individuals instructional level should be used.

3. Evidence that the DRP is culture fair. Using the data collected for the aforementioned information acquisition and gain study, a set of demanding tests for bias were performed. The results of these analyses of variance showed that DRP ability scores do not have to be conditioned by knowledge of either sex, race, or socioeconomic status in making predictions concerning what students will do on prose reading tasks external to the DRP.

4. The measurement of individual gain. Evidence was obtained that DRP tests can measure individual gain over a five month interval even when broad-band tests forms (covering a wide range of readability) are used. Two randomly constituted "parallel" broad-band forms of the DRP were administered to fourth and sixth grade students, with a five-month interval between tests. All ability levels evidenced gain, and this gain was statistically significant for all but the highest ability group in grade 4.

5. Reliability: Making the same decision concerning ability with repeated measurements. When randomly constituted parallel forms of the DRP were administered with only a two-week interval between tests, no real differences were observed between ability scores. Thus the DRP does in fact make the same decision concerning ability with repeated measurements. This measure of reliability, unlike True Test Score theory correlations, does not depend upon comparisons among students. Instead, it measures within-

individual performance over a time period when no change in ability is expected. In a similar fashion, the test for gain (see above) determines whether within-individual change is obtained over a period of time in which change *is* expected. Both tests are in accord with the constructs of reliability and change as they are elaborated in the natural sciences.

III. Discussion

The DRP's unique characteristics make it a more sensitive measure of individual gain in reading ability than has heretofore been available. The methods of test construction used for the DRP serve to minimize such artifacts of measurement as regression effects, which often thwart efforts to assess individual growth. Thus, the DRP could be used to evaluate more sensitively the outcomes of reading instruction or gauge an individual's growth over the school year.[10] It is the first reading test to measure individual gain reliably.

On a larger scale, the methodology used to develop the DRP can be used to set rational standards against which to measure growth. Three important questions must be answered in order to set such standards. What reading ability must an individual have to be a literate adult? How much practice and consolidation must take place with materials at a given level of difficulty before an individual should be given materials at the next reading level? Are there individual growth curves? Since none of these questions can be answered precisely, given the present state of knowledge, the setting of rational standards is a future goal. However, it is presently possible to examine *de facto* standards. The readability of assigned materials that students are presumably expected to master during a given school year can be determined. The DRP can then be used to assess whether these *de facto* standards are being met and/or to revise and individualize the standards.

The DRP promises to have an impact on the classroom and the individual student in other ways. The DRP can be used to match instructional materials to the student's reading capabilities. This valuable function would help to prevent the boredom and frustration that may occur when students are faced with materials that are too easy or too difficult for them, and would tend to maximize the effective use of classroom time. Being able to match materials with student abilities objectively should also be a great aid when making decisions concerning the purchase of text books, particularly when individualizing instruction. In addition, the DRP could be used for placement purposes: for grouping students of comparable ability, or for deciding if a particular student is capable of reading at the level which an examination of *de facto* standards reveals is required of students in a given grade.

Footnotes

[1] Some publishers simply present their tests, with no attempt at all to describe what is being measured; other publishers name and briefly describe, or simply name, subskills that their tests are supposed to be measuring. However no publisher (to our knowledge) operationally defines the construct "reading comprehension" in terms of specifying the procedures or operations by which it can be measured.

[2] See Stanley (1971) for a detailed discussion of the dilemmas associated with obtaining reliable difference scores on normed tests.

[3] It is not a simple task to identify, enumerate and validate the component skills involved in reading with comprehension. First, there is appreciable empirical evidence from psychometric studies that all the reading comprehension subskills (excluding perhaps vocabulary) can be accounted for by a single factor (e.g., Thurstone, 1946; Hunt, 1957; Bormuth, 1969b). Second, linguistic and psycholinguistic analyses and empirical studies show that (by comparison with emerging results from an information-processing point of view) the typical skills measured on either norm-referenced or criterion-referenced tests are invalid nominal factors (e.g., Cooper and Petrosky, 1976; Farr, 1969). In other words, the formulation of a set of component factors (or skills) in reading with comprehension is a state-of-the-art problem.

[4] These items are rooted in a skills view of reading. However, as noted earlier, the evidence for the validity of the various subskills as separable, measurable components of the reading process is weak, and it is not known how the subskills (if they are valid) combine to comprise reading with comprehension. In addition to problems in the theory underlying the items, the "typical" items often suffer from technical deficiences of the sort noted earlier: students can answer the items without reading the prose.

[5] Readability is a function both of the difficulty of the material and the ability of the reader. It can be partitioned into stimulus difficulty and ability. In our discussion we are concerned with the difficulty inhering in the stimulus materials, and shall therefore speak of prose (or text) difficulty.

[6] For example, Fry (1968) reports intercorrelations among five readability measures ranging from r = .64-.98, with all but three of the values in the intercorrelation matrix above r = .80 and most above r = .90. In earlier (Koslin, et al., 1975) a correlation of r = .90 was found between Dale-Chall (1948) and Bormuth (1969a) readability scores.

[7] In doing so, we are setting aside for a while questions concerning the validity of the current operational standards. The expectations or *de facto* reading standards (may be assumed to) derive mainly from the opinion of curriculum experts concerning what students ought to be able to achieve in different grades. These standards probably also reflect, at least in

part, the opinions of editors and publishers concerning grade level placement of materials; these opinions are undoubtedly influenced by the application of readability formulas that are inappropriately interpreted as prescribing what students ought to do in a given grade. The formulation of more rational grade level standards must remain a task for the future, as it will require information concerning how adult competence standards are to be defined as well as increased knowledge about developmental learning curves for reaching those standards.

[8] Note that the Bormuth mean cloze scores form a scale where high scores indicate easy materials, while Rasch values increase with difficulty as do, by definition, high mean and median passage difficulties; hence the negative correlation.

[9] Scores for the lowest ability groups -- those scoring essentially at the chance level for random guesses on a multiple choice test -- did increase over the two-week interval. However, since this regression is best interpreted as being due to guessing, it is ignored for purposes of interpreting the remainder of the data from this study.

[10] Note, though, that individual gain cannot be measured reliably for those students who get very few items correct on broad-band DRP tests, i.e., for those students who are essentially non-readers. DRP tests centered on ability for these students would be prerequisite to measuring gain for them.

References

Adult Performance Level Project Staff. The adult performance level study. Austin: University of Texas, 1973.

Anderson, S.B., Ball, S., Murphy, R.T., & Associates. Encyclopedia of educational evaluation. San Francisco: Jossey-Bass, 1973.

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Bormuth, J.R. Readability: a new approach. Reading Research Quarterly, 1966, 1, 79-132.

Bormuth, J.R. Cloze test readability: Criterion reference scores. Journal of Educational Measurement, 1968, 5, 189-196.

Bormuth, J.R. Development of readability analyses. Final report, Project No. 7-0052, Contract No. OEG-3-7-070052-0326, Office of Education, Bureau of Research, U.S. Department of Health, Education, and Welfare, March, 1969. (a)

Bormuth, J.R. Factor validity of cloze tests as measures of reading comprehension ability. Reading Research Quarterly, 1969, 4, 358-365. (b)

Bormuth, J.R. Development of standards of readability: Toward a rational criterion of passage performance. Final Report, Project No. 9-0237, Grant No. OEG-0-9-230237-4125, Bureau of Research, Office of Education, U.S. Department of Health, Education, and Welfare, 1971.

Brinton, J.E. & Danielson, W.A. A factor analysis of language elements affecting readability. *Journalism Quarterly*, 1958, *35*, 420-426.

Campbell, D.T. & Fisk, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Carroll, J.B. Word frequency studies and the lognormal distribution. In E.M. Zale (Ed.), *Proceedings of the conference on language and language behavior*. New York: Appleton-Century-Crofts, 1968.

Carroll, J.B., Davies, P., & Richman, B. *The American Heritage word frequency book*. Boston and New York: Houghton Mifflin Co. and American Heritage Publishing Co., 1971.

Carver, R.P. Reading tests in 1970 versus 1980: Psychometric versus edumetric. *Reading Teacher*, 1972, *26*, 229-302.

Chall, J.S. *Learning to read*: *The great debate*. New York: McGraw-Hill, 1967.

Cleary, T.A., Humphreys, L.G., Kendrick, S.A., & Wesman, A. Educational uses of tests with disadvantaged students. *American Psychologist*, 1975, *30*, 15-41.

Coleman, E.B. *On understanding prose*: *some determiners of its complexity*. Washington, D.C.: National Science Foundation, 1965. (NSF Final Report GB 2604)

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J.H., Mood, A.M., Weinfeld, F.D., & York, R.L. *Equality of educational opportunity*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, U.S. Government Printing Office, 1966.

Cooper, C.R. & Petrofsky, A.R. A psycholinguistic view of the fluent reading process. *Journal of Reading*, 1976, *20*, 184-207.

CTB/McGraw-Hill. *Examiner's Manual for the California Achievement Tests Level 3, Form A*. Monterey, Calif.: CTB/McGraw-Hill, 1970.

Dale, E. & Chall, J. A formula for predicting readability. *Educational Research Bulletin*, 1948, *27*(1), 11-20, 28; *27*(2), 11-28.

Davis, F.B. Research in comprehension in reading. *Reading Research Quarterly*, 1968, *3*, 499-544.

English, H.B. & English, A.C. *A comprehensive dictionary of psychological and psychoanalytical terms*. New York: David McKay Company, 1958.

Farr, R. *Reading*: *what can be measured*? Newark, Del.: International Reading Association, 1969.

Flesch, R. A new readability yardstick. *Journal of Applied Psychology*, 1948, *32*, 221-233.

Frederiksen, J. Personal communication, 1975.

Frederiksen, J. *School effectiveness and equality of education opportunity*. Report submitted to the Carnegie Corporation of New York, December, 1975 (b).

Frederiksen, N. *Some emerging trends in testing*. Paper presented at the Summer Study on Research in Testing, Falmouth, Mass., August, 1978.

Fry, E.B. A readability formula that saves time. *Journal of Reading*, 1968, *11*, 513-516, 575-578.

Garner, W.R. *Uncertainty and structure as psychological concepts*. New York: John Wiley & Sons, 1962.

Glaser, R. & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1971.

Gray, W.S. & Leary, B.E. What makes a book readable: an initial study. Chicago, University of Chicago Press, 1935.

Greene, F.P. A modified cloze procedure for assessing adult reading comprehension. Unpublished doctoral dissertation, University of Michigan, 1964.

Hackman, R.C. & Kershner, A.M. The determination of criteria of readability. University Park, Md.: University of Maryland, 1951. (Technical report for Office of Naval Research Contract NR 153-024)

Harris, L. & Associates. The 1971 national reading difficulty index: a study of functional reading ability in the U.S. for the National Reading Center. New York: Harris & Associates, 1971.

Hayes, J.D. A grade placement study of primary reading material. Unpublished master's thesis, University of Southern California, 1934.

Herdan, G. Type-token mathematics. The Hague: Mouton, 1960.

Howes, D.H., & Solomon, R.L. Visual duration threshold as a function of word probability. Journal of Experimental Psychology, 1951, 41, 401-410.

Hunt, L.C. Can we measure specific factors associated with reading comprehension? Journal of Educational Research, 1957, 51, 161-172.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. Inequality: A reassessment of the effect of family and schooling in American. New York: Basic Books, 1972.

Klare, G.R. The measurement of readability. Ames: The Iowa State University Press, 1963.

Koslin, S., Koslin, B.L., Pargament, R., & Pendelton, S. An evaluation of fifth grade reading programs in ten New York City Community School Districts, 1973-1974. New York: Riverside Research Institute, 1975.

Koslin, S., Koslin, B.L., Langley, C., & Josephson, B. Design concepts for a measure of effectiveness in reading. New York: Riverside Research Institute, 1973.

Kyte, G.C. Calibrating reading material. Elementary School Journal, 1925, 25, 533-546.

Lorge, I. Predicting reading difficulty of selections for children. Elementary English Review, 1939, 16, 229-233.

Lorge, I. The Lorge and Flesch readability formulas: a correction. School and Society, 1948, 67, 141-142.

MacGinitie, W.H. & Tretiak, R. Sentence depth measures as predictors of reading difficulty. Reading Research Quarterly, 1971, 6, 364-377.

Mathews, C.O. The grade placement of curriculum materials in the social studies. Columbia University Contributions to Education, No. 241. New York: Teachers College, Columbia University, 1926.

McCall, W.A. & Crabbs, L.M. Standard test lessons in reading. New York: Bureau of Publications, Teachers College, Columbia University, 1925.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Mosteller, F., & Moynihan, D.P. (Eds.) On equality of educational opportunity. New York: Random House, 1972.

Murphy, R.T. Adult functional reading study. Final Report, Project No. 0-9004, Grant No. OEC 0-70-4791 (508). Princeton: Educational Testing Service, 1973.

Pierce, J.R. & Karlin, J.E. Reading rates and the information rate of a human channel. Bell System Technical Journal, 1957, 36, 497-516.

Postman, L. & Rosenzweig, M.R. Perceptual recognition of words. Journal of Speech Disorders, 1957, 22, 245-253.

Potter, T.C. A taxonomy of cloze research, Part I: Readability and reading comprehension. Los Alamitas, Ca.: Southwest Regional Laboratory for Educational Research and Development, 1968.

Rankin, E.F., Jr. The cloze procedure -- its validity and utility. In O.S. Causey and W. Eller (Eds.), Yearbook of the National Reading Conference, 1959, 8, 131-144.

Rankin, E.F., Jr. The cloze procedure -- a survey of research. In E.L. Thurston and L.E. Hafner (Eds.), Yearbook of the National Reading Conference, 1965, 14, 133-150.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Studies in Mathematical Psychology, I. Danmarks paedagogiske Institut, 1960.

Shannon, C.E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27, 379-423, 623-656.

Skinner, B.F. Verbal behavior. New York: Appleton-Century-Crofts, 1957.

Slinde, J.A. & Linn, R.L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Smith, F. Psycholinguistics and reading. New York: Holt, Rinehart, and Winston, 1973.

Smith, F. Comprehension and learning. New York: Holt, Rinehart, and Winston, 1975.

Smith, H.E. The validity of teachers' judgments of difficulty in curricular materials. Journal of Educational Psychology, 1930, 21, 460-466.

Spache, G. A new readability formula for primary grade reading materials. Elementary School Journal, 1953, 53, 410-413.

Staiger, R.C. Certain language factors in the readability of primary reading textbooks. Journal of Educational Research, 1955, 48, 589-596.

Stanley, J. Reliability. In R.L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

State Education Department. The further development of an effectiveness measure in reading. Albany: The University of the State of New York, The State Education Department, 1975.

State Education Department. The further development of an effectiveness measure in reading: Phase II. Albany: The University of the State of New York, 1977.

Stevens, S.S. Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), Handbook of experimental psychology. New York: John Wiley & Sons, 1951.

Stolurow, L.M. & Newman, J.R. A factorial analysis of objective features of printed language presumably related to reading difficulty. Journal of Educational Research, 1959, 52, 243-251.

Taylor, W.L. "Cloze procedure": a new tool for measuring readability. Journalism Quarterly, 1953, 30, 415-433.

Taylor, W.L. Recent developments in the use of "cloze procedure." Journalism Quarterly, 1956, 33, 42-48.

Thorndike, E.L. & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College. 1944.

Tinkelman, S.N. Planning the objective test. In R.L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Thurstone, L.L. Note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 185-188.

Tuinman, J.J. Thorndike revisited -- some facts. Reading Research Quarterly, 1971, 7(#1), 195-202.

Tuinman, J.J. Determining the passage dependency of comprehension questions on 5 major tests. Reading Research Quarterly, 1974, 9, 206-223.

University of the State of New York. Report of the Commissioner to the Governor and Legislature. Albany: The State Education Department, 1974.

Washburne, C. & Vogel, M. What books fit what children? Schools and Society, 1926, 23, 22-24.

Wainer, H. & Pratscher, G. A scheme for the robust estimation of ability in a latent trait model (working paper). Elmsford, N.Y.: Touchstone Applied Science Associates, 1976.

Wright, B.D. Solving measurement problems with the Rasch Model. Journal of Educational Measurement, 1977, 14(2), 97-117.

Wright, B.D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Zipf, G.K. The psycho-biology of language. Boston: Houghton Mifflin, 1935.

PART THREE: COMMITTEE REPORTS

Testing to Improve Teaching and Learning: The View from Cognitive Science 337

Teaching, Testing, and Learning: The Approach Through Research on Classroom Processes 357

Hard Decisions from Soft Data 375

TESTING TO IMPROVE TEACHING AND LEARNING: THE VIEW FROM COGNITIVE SCIENCE

Report of Committee 1

Allan Collins
Robert B. Davis
D. Ross Green
Leon Henkin
David Klahr, Chair
George S. Miller
Amado M. Padilla

CONTENTS

I. INTRODUCTION 338
- A. Relation of Tests to the Purposes of Learning and Teaching 338
- B. A Vision of a New Learning and Testing Environment 342

II. PROPERTIES AND USES OF TESTS 343

III. PERCEIVED COSTS AND BENEFITS OF TESTING 344

IV. DESIDERATA FOR NEW TEACHING, LEARNING AND TESTING ENVIRONMENTS 346
- A. Tests Involved in Integrated Classroom Activity 346
- B. Accountability of the Testing Procedure 347
- C. Motivating Aspects of Testing 347
- D. Educational Value of Feedback Testing Procedures 347
- E. Test Diagnosis and Individual Differences 348
- F. Validity of the Tests 348
- G. Relation Between Form and Purpose of Tests 348
- H. Tests Which Assess Different Kinds of Problems 349

V. RESEARCH AND DEVELOPMENT 349
- A. Development of Activities Integrating Testing and Teaching 350
- B. Detailed Descriptions of Student Processing Strategies 351
- C. New Data Extraction Analysis and Aggregation Procedures ... 352
- D. Studies of Children's Learning and Thinking Processes 353
- E. Computer Security and Privacy: What to Record, What to Keep 353
- F. Research on Teacher's Understanding of Student's Knowledge 353
- G. Research on Testing Students in Groups 353
- H. Factors and Biases in Individual Decision Making 354
- I. Test Validation 354
- J. Research on Current Uses of Tests 354

REFERENCES 356

I. INTRODUCTION

In this section we give a brief summary of the way that testing is related to teaching and learning, and we suggest some immediate improvements to the testing process. Then we sketch a future vision of an idealized testing situation. The rest of the report provides an elaboration and further specification of the issues raised in this introduction.

A. Relation of Tests to the Purposes of Learning and Teaching

There are many reasons for testing students, not all of which are related to the purposes of learning and teaching. If we set aside, for the present discussion, those reasons that arise from administrative or political needs, we are left with the need to inform students of their progress in learning, and to inform teachers of what should be taught. The most direct relation of testing to learning and teaching is feedback to the student and the teacher. Let us, in order to have a neutral term, refer to tests administered for such purposes as "feedback tests."

And let us, in order to have a well defined goal, consider whether feedback testing could be performed more effectively or more efficiently than it is in most classrooms at the present time. It may well turn out that the results of improved feedback tests could still serve other purposes, or it could turn out that tests optimized for feedback to the student and teacher would have to be supplemented by other tests for other purposes. For the present, however, let us take the feedback of information as the testing function to be optimized. Then we can explore that challenge in its own right, independently of other uses of tests. If it should turn out that there is a tradeoff -- that improved feedback tests result in poorer administrative information, for example -- that would be an important conclusion for those responsible for the future of educational testing in this country.

A characteristic, if not defining, property of feedback tests is that the student stands to benefit from taking them. In order to optimize them, therefore, we should consider first of all the benefits that they confer. If those benefits can be strengthened, feedback testing should be improved thereby.

For the student, the most direct consequence of taking a feedback test is a grade -- a number or alphabetic equivalent that gives the student some notion of where he or she stands in comparison to others in the same life situation. It is generally assumed that such information is beneficial to the student. The assumption is usually supported by some version of the psychological theory of reinforcement, even when the tests are designed to give discouraging feedback to 50% of the students.

It is also generally assumed that there is an indirect benefit to the student that is mediated through the teacher. The results of the tests are imagined to give the teacher guidance as to what the students do not know, and so need to be taught. The student benefits by receiving instruction in what he or she still needs to learn. This view is defensible only when the content of the tests and the content of instruction are closely related. Too often that relation is achieved only by tailoring the instruction to the requirements of the test, but even when the teacher designs the test to provide information about that teacher's goals, the results characteristically generate a wide distribution of scores: different students need to be taught different things. What is to be done as a consequence of the feedback provided by the test is seldom clear.

Recognition of the problems associated with feedback testing has led to a call for more diagnostic tests. As this is usually understood, the purpose of the test is not to provide some overall figure of merit, but to provide a variety of measures for different aspects of the skills or content that is to be learned. The student learns, not merely that he or she is above, at, or below average in arithmetic, for example, but that, say, long division is in good shape and more work on fractions is needed. Diagnostic testing is impossible without a close association between the content of the test and the content of instruction, but serious attempts to provide diagnostic feedback have, by and large, been discouraging. They do not motivate a student to make up the deficiencies, and they do not clarify what a teacher should do when faced with evidence of student diversity. The constraint to provide more information, however, does mean that more time must be spent in testing -- and every hour spent testing is an hour less spent learning and teaching.

Diagnostic feedback makes sense only when coupled with individualized instruction. It is beyond the mandate of this committee to evaluate the cost-effectiveness of individualized instruction, especially when so many different techniques for individualization are currently in competition. It is generally recognized, however, that the demands such systems place on the teacher are very different from those most teachers are trained to meet, and that the social consequences of individualization for the student are often undesirable.

One great virtue of individualized instruction is that testing is so closely integrated with the goals of learning and teaching. Indeed, in some forms the process of instruction is reduced entirely to repetitions of the cycle: teacher's question, student's answer, teacher's answer. The student can compare his own answer immediately with the "correct" answer; the teacher can use the student's answer to select the next question; the whole transaction is effectively a test of the student. Testing, learning, and teaching are all integrated into a single instructional process.

This method of instruction resembles a feedback test even more when the teacher is replaced by a printed form. When the entire process is programmed in advance, however, many of the advantages of individualization are lost. Flexibility can be introduced by computers, which can provide alternative questions as a function of the student's response, but the cost is usually that the only form of question available is the multiple-choice form.

Considerable ingenuity and effort have been invested in these developments, usually under the banner of better teaching. Without attempting to determine whether that goal has actually been achieved, we can ask a different question. Consider the claim that such programs of instruction provide better feedback testing. The support for such a claim would be that, viewed as tests, such programs are tests that teach as well as test.

The fact that such feedback testing that teaches is already available suggests a new challenge for this growing technology. Suppose that we were to insist that every feedback test should carry instructional benefits for the student. That is to say, not only would a student benefit from knowing where he or she stood, but would also benefit from knowing more about the subject matter at the end of the test.

Tests are often constructed with the deliberate intent of preventing such learning. That is to say, it has been considered desirable if a test could be administered twice with the same results -- such tests are said to be reliable. A test that teaches, therefore, necessarily abandons this conception of reliability of measurement as one of its goals. In its place, one would set the goal that the student should score perfectly on a second administration of the test. In a sense, a test that teaches should self-destruct.

Tests that teach are a special class of feedback tests, and the challenge to develop such tests need not be locked into a cycle of questions and answers. To take a very different example, consider how a person learns to play golf. Each time the ball is hit toward a target, the flight of the ball provides immediate feedback on the success of the swing. Each time a round is played, the player's score tells him where he stands in his efforts to master the game. A difficult course is often said to be "a good test" of the player's skills. In such cases, testing and learning are so intimately related that it requires a special frame of mind to distinguish them.

When we adopt that frame of mind, however, we can recognize a variety of "tests" that teach. Games provide paradigm instances, on three counts: (1) motor responses provide their own feedback; (2) competitors are quick to take advantage of mistakes; (3) wins and losses -- who can beat whom -- tell the player where he or she stands. Their shortcoming, of course, is that playing a game teaches a person only how to play that game; what is learned, if not trivial, is at least unrelated to most academic subject matter.

Academic subjects can be presented in the form of games, as educators learned long ago -- the spelling bee provides a familiar example. More contemporary examples offer a student a game played against a computer, where success demands mastery of some academic skill, e.g., knowledge of fractions might be required in order to aim a dart accurately at targets on the oscilloscope, or the ability to read might be required in order to follow instructions. The game format is a rich source of examples that combine testing and learning.

The game of twenty questions suggests a reversal of the usual roles: the student can ask the questions, instead of the teacher. The quality of the question asked can be scored, and the number of questions required to determine the answer gives feedback to the student. Indeed, cognitive psychologists have found interesting developments in the kinds of questions children ask at different ages. Such procedures might be particularly appropriate for teaching categorizations, e.g., of animals, minerals, or vegetables, but any problem that can be solved by successive approximations could also serve as subject matter.

The impression that a large domain of tests that teach is waiting to be explored by imaginative educators is fed by such devices as BUGGY, which is capable of diagnosing a systematic error by the student (a "bug" in the student's program). BUGGY reveals precisely what the student needs to be told; more advanced students can pose deliberate problems for the machine to solve, or can try their hand at finding a bug selected by the machine. BUGGY is one of the most sophisticated testing-teaching systems currently available.

BUGGY makes it clear, however, that sophisticated developments of tests that teach must be based on a very detailed knowledge of the subject matter. The subject of arithmetic operations is well enough known, and the operations that students perform in trying to compute arithmetic answers are sufficiently limited, that it is possible to realize such a system on the basis of current knowledge. It would be much more difficult, in the current state of cognitive science, to devise such a system for diagnosing student errors in reading comprehension. It might provide an interesting challenge to cognitive scientists to analyze the mental processes involved in various cognitive skills in sufficient detail to support a BUGGY type of testing-teaching system.

If this should prove to be direction of future test development, it is likely that we will conclude that optimal feedback tests are not optimal for the purposes of educational management and administration. Viewed as tests, such systems would give as scores something that might be interpreted as the student's current location in a large, abstract, multidimensional representation of the subject matter. Such information would have to be grossly degraded in order to conform to a requirement that students be linearly ordered with respect to some norm or criterion.

In any case, it seem clear that feedback tests currently in use in most schools are not optimal, and that recent developments in computer technology, test theory, and cognitive science have made possible a much more productive use of the time students spend taking tests. The future should see a much clearer distinction between testing instruments that assess what a student has already achieved and those that assess what a student is prepared to learn next.

B. A Vision of a New Learning and Testing Environment

In our view of the future school tests, as we know them, would cease to exist. The intrusive, specialized, institutionalized activity called test-taking would be absorbed into a new form of learning and testing environment. Computers could accept inputs from students and teachers on an almost continual basis, extracted from the rich tapestry of ongoing learning activities. Instructional systems would accumulate an educational portfolio including a wide range of interrelated performance and situational descriptions. We would be as unlikely to cease all instructional activities in order to "test" a student as we are to stop conversing with a child in order to "test" his or her linguistic competence. Instead, a test would be a particular aggregation and analysis procedure applied to a continuously collected data base. Some of these aggregations would be utilized to have an immediate impact on ongoing activities, while others would be quite remote from the moment of data collection.

This vision has two central problems: one concerns feasibility, the other concerns potential abuse. These are complex and interrelated issues. Before turning to them, however, we want to exemplify our testing vision with an analogy.

Consider the evaluation of ability, knowledge, and skill in team sports. Although a professional ball player rarely takes a batting test or a running test or a fielding test, we can find endless tabulations of descriptive statistics that inform us about his quality on many dimensions. These statistics are based on observations from the ongoing course of play. Immediate aggregation is used for on-line decisions (e.g., a pitcher has walked 15 men in a row: take the bum out), while after-the-fact aggregation produces an incredible array of constrained conditional descriptions (e.g., number of fouls hit against left-handed pitchers). In the case of football, videotape records are maintained which show players' performance on every play in every game. Such tapes are used both to assess and control game-time performance as well as to derive qualitative features of players and associated diagnostic and remedial procedures that may not have been anticipated at the time of initial observation.

We are well aware that Public School 258 is not identical to the Los Angeles Rams, but the analogy is intended to exemplify a current use of a non-testing situation in which a record of continuous observation provides a basis for both immediate and remote evaluation and decision making.

Can we do this in the schools? Do we have the technology, the knowledge, and wisdom to do it now? It is certainly not possible now in its full form, but some of the pieces already exist. Whether or not this vision is feasible in the near future depends in part on whether sufficient resources are allocated to the necessary research in the cognitive sciences and in the development of information processing technology. These issues are addressed in the following sections.

Should we do it, if we could? How can we ensure that our benign vision, aimed at improving the sensitivity, responsiveness, and relevance of our efforts to help children learn does not lead to the inherent dangers of extensive record keeping? How do we maintain accountability, privacy, individual choice? How do we avoid undesirable social consequences that potentially--but not necessarily--result from a sophisticated record-keeping technology? We believe that these questions, when raised at the very start of a research and development effort, can significantly contribute to the design of such systems. Many of the items on the research and development agenda below reflect our concern with the social-policy implications of the system we envision.

II. PROPERTIES AND USES OF TESTS

Testing is not the same as observing, although of course observation is an important component of testing. What else is involved? It seems to us that there are two other essential components in the kind of tests here considered. First, what is observed is the reaction of the student to a situation that has been designed to indicate degree of learning, command of skill, extent of knowledge, and depth of understanding, of a specific subject area that is treated in an instructional program. Second, the observation is focused on those aspects of the student's reaction that are considered relevant to the information sought by the teacher. Indeed, in many cases the teacher is prepared in advance of the test-observation to record the outcome in a specific way that perhaps limits what is observed, but is intended to help the teacher help the student to advance in the instructional program.

How do tests function? They help various individuals to make judgements--primarily about the student tested, but sometimes about the teacher, the school, the course, or the study materials. Here we are primarily interested in the teacher's judgements, about the student, and the student's judgements about himself. Almost always, the teacher is judging in order to anticipate the future behavior of the student or to understand the student's current difficulties. How will the student perform in next week's lesson? Will the student be able to utilize effectively the term's classroom experience when he goes on to college, or in a job related to the subject of the class? Will the student remember what was taught when challenged by appropriate circumstances? Is the student sufficiently confident about his mastery of the subject so that he will venture to undertake a creative project for which the mastery seems necessary? These are some of the kinds of questions about the student's future behavior, for which the teacher looks for help from tests.

It is obvious, however that for reliable interpretations to flow from test results, it is necessary to fit these test-observations into some sort of model, or theoretical framework. Without such a theory, a teacher is forced to rely on raw intuition, supplemented on occasion by statistical correlations, in order to pass from observations and test scores to the inferential judgments that are relevant; unfortunately, such reliance too often leads the teacher to incorrect conclusions.

The theory that has, explicitly or implicitly, dominated most thinking about tests is that they provide measures, analogous to length or weight, that can be used to arrange children, and what they have learned, in a simple linear order. This theory supported initial research on testing, but it is obviously inadequate to our present conceptions of the complex nature of human knowledge.

A giant step in improvement of testing would stem from development of better theory--we expect this giant step to take place but not for a while. We discuss conditions that might foster such development in the last parts of this report.

Smaller but important steps in improvement might take place from use of existing theory and technology. We should recall that the essential form and structure of educational testing has not changed in over fifty years. Surely we have learned something since the days when Babe Ruth was playing in Yankee Stadium. Computers, for example, give us the capacity to store and retrieve incredible quantities of information--no longer do we reduce a student's achievements to a single number. We also know that simulations increase in validity as they move to closer and closer approximations of reality--why then, shouldn't we measure a student's capacity to learn something while the student is actually learning? Such small but important steps are what this report is about.

III. PERCEIVED COSTS AND BENEFITS OF TESTING

As with any other human enterprise, testing has generated a series of perceptions and concerns about both its costs and its benefits. Some of these are agreed upon, others are subject to dispute.

While these issues arise most often in discussions of the use of published or otherwise externally developed tests, teacher-made tests are frequently similar in certain of these respects. The prevalence and seriousness of most of these problems has not been adequately investigated. Nevertheless the frequency with which such concerns are stated indicates a need to deal with them seriously, regardless of the degree to which these perceptions of reality are accurate.

Perceived benefits of testing:

1. Testing is useful in assisting both the teacher and the learner in the educational process since it provides information on how each is doing.

2. Testing can provide diagnostic information concerning where the learner is having difficulty so that instruction can be provided, and where he has strengths upon which to base further learning.

3. Testing is a necessary process if wise decisions are to be made about who to educate, for how long, and with what subject matter.

4. Testing can be carried out fairly so that equitable decisions about education can be made.

5. Testing is carried out according to scientifically sound precepts which ensure that tests are appropriate to their intended purpose.

Perceived costs of testing:

1. Resource Consumption -- Testing often diverts the student and the teacher from instruction and learning *per se*. It is believed that in many schools too much student time and too much money are spent on testing.

2. Intrusive and Obtrusive -- Testing forms such a major part of the school environment that for some children -- and their teachers -- tests become pervasive and threatening, and often isolate the child from his peers and teacher. This becomes especially important when it becomes clear to the child that some decision that can affect him/her adversely will be based on the test.

   It is reasonable to assume that teacher attitudes towards any particular test play a major role in these student reactions, but whether students feel threatened because teachers do is not known. Too little is known about the conditions and circumstances surrounding test administration to judge how often tests seriously interfere with learning and instruction. Clearly they should never do so.

3. Irrelevance -- Testing is often poorly integrated with teaching and learning, and even less connected to performance in "real" life activity. This concern extends the argument to suggest that testing should have some basis in the types of skills and knowledge the child needs in his/her environment. This may be most important when the "real" life application is not apparent to the child. Another way of stating this is that tests must be relevant to some subject matter and the child must see the relevance.

4. Insensitivity -- Testing is purportedly insensitive in varying degrees to the following:

   a. Process -- Tests are so constructed that right/correct answers are the only responses that the tests seek. There is no place for analyzing the process by which the child arrived at correct or incorrect response. Often the process is more informative than the final answer.

   b. Linguistic differences -- Tests assume a common linguistic core/level for all children. For both cultural and/or socioeconomic reasons, there may be differences in linguistic competence in English between students in the same classroom, and across schools and geographic regions. Such differences in linguistic competence may place some children at a disadvantage.

Although the degree to which cultural differences among students affect test performance is not known, it is clear that there is some effect on most tests, published and teacher-made alike. Procedures for dealing with this sort of problem appear to be less widely known and/or employed than is desirable and reasonable.

c. Experiential differences -- Tests assume a common experiential level for all children. Again, for both cultural and/or socioeconomic reasons there may be differences between students in the same classroom, and across schools and geographic regions in experiences, familiarity, and knowledge required by some test items or tests. Such differences may place some children at an unfair disadvantage.

d. Motivational differences -- Testing assumes that all students will be more or less equally motivated when taking a test. Individual fluctuations in a child's motivation due to illness, interest, stress, etc. are built into testing usually. For example, if a child doesn't feel like taking a test today, for whatever reason, it is unlikely that the test can be put off until tomorrow.

e. Knowledge -- Often tests don't measure what a child knows, but what he/she doesn't know. In some cases this may be desirable, but in other cases, may be harmful to the child's self concept as a learner.

IV. DESIDERATA FOR NEW TEACHING, LEARNING, AND TESTING ENVIRONMENTS

We think it is possible to create new kinds of testing procedures that would alleviate some of the controversies that surround current testing practices. These can be embedded in the activity-based learning environment described in the Introduction to this section. There are a number of specific objectives that should be met in designing testing procedures for such an environment.

A. Tests Involved in Integrated Classroom Activity

Testing should function like the teacher who continuously monitors the student's behavior while the student carries out various learning activities. These may be paper and pencil activities or they may be computer-based activities. In either case the testing procedure should aggregate over extensive data derived from the student's performance, including teacher observations.

Ideally, the measurements should be non-interfering. They should be derived from the traces the student leaves behind in carrying out different learning activities. This enables the testing to be both non-threatening and non-time-consuming, because it is embedded in the student's daily classroom activities. This also realizes the objective that the testing procedure diverge as little as possible from the student's normal learning experiences.

B. Accountability of the Testing Procedure

There are a number of dangers inherent in any testing procedure that aggregates extensive data on the student. These records can potentially be used as a kind of dossier from which the student cannot escape during his academic career or beyond. Misinformation can creep into the record. Therefore, there should be safeguards built into the use of testing records to protect students and teachers.

The major safeguards we propose are:

- Each piece of information added to the record should have a prespecified lifetime after which it is destroyed.
- The accomplishments in the student's record should become the student's property. This he can treat as a kind of resume, from which he can delete any material he desires.
- Parents should be provided easy access to the kinds of information that would assist them in their instructional role outside the classroom.
- The administration, construction, and utilization of the testing procedures should be clearly understood by students, parents, teachers and any other interested parties.
- The locus of decisions about test content and use should be clear.

C. Motivating Aspects of Testing

The tests should be constructed around activities that are interesting to students, so that they care about performing well. The situation should be non-threatening to the student. The objectives of the testing should be clear to the students in order to gain their cooperation.

D. Educational Value of Feedback Testing Procedures

The tests should be constructed to have educational value for students. To this end the tests should provide useful feedback to the student; both as to what he has accomplished, what he is doing wrong, and what remains to be tackled. Ideally, feedback testing should occur at the edge of what the student is learning so that it can inform further teaching and choice of educational activities. But we should also test whether the student is relating recent instruction to more distant topics.

Every teacher has an obligation to help his/her students toward independent learning, for as students progress through the elementary and secondary grades, and toward the post-school years, they become increasingly responsible for their own learning. Obviously self-learning will be most effective if it incorporates self-evaluation -- assessment of progress, and of strengths and weaknesses. It follows that tests should be developed to help students with self-evaluation, and that teachers should help students learn how to use such tests. These tests should bear not only on current topics of the classroom curriculum, but should help the student to adjust his perception of the pattern of his abilities in the light of new learning experience.

E. Test Diagnosis and Individual Differences

It is universally recognized that the performance of pupils is significantly affected by a variety of individual differences. Students normally come into a given class from different earlier classes -- often different schools, even different parts of the country; they come from homes with different attitudes, different study conditions; they develop with different learning styles. Because these differences affect test outcomes as well as the ongoing process of learning, it is necessary to take them into account in analyzing and reporting test results.

F. Validity of the Tests

Tests must be constructed so that they fairly represent the student's ability in those tasks in the real world or school environment, where the skills measured will be required (i.e., "ecological" validity). Not only must measures collected predict performance on representative tasks, but also the learning activities in which the measures are embedded must be shown to teach skills that are genuinely applicable in real world settings.

A second kind of validity it is important to establish is how well the test measures a student's performance over the entire range of student behavior. Very often tests are valid in the middle range of behavior, but misrepresent the student's skills in the extremes of the distribution. This can lead to misdiagnosis of students' problems or misallocation of teaching efforts.

G. Relation Between Form and Purpose of Tests

Just as tests deal with many diverse aspects of the learning process, so their formats can vary widely. While the familiar multiple-choice style of questions may be appropriate to checking the content of memory, the ability to discriminate shades of meaning, or to find numerical answers to problems, quite different forms of testing are appropriate if one is seeking to evaluate creative power, or to ascertain what methods of problem-solving a student is using or can design. Of course, the more open-ended a test is, the more difficult it becomes to assess the responses it evokes. The design of a test includes not only the questions presented to the student, but the way to deal with his response. There is a need, then, for tests that will deal effectively with both eliciting and assessing each of the many aspects of learning.

H. Tests Which Assess Different Kinds of Problems

There are a number of aspects of student's performance that any testing procedure should attempt to assess in addition to correct performance. We will briefly describe several of these:

- Students' confusion -- A teacher can often tell if a student is confused by a puzzled look on the face. Similarly, testing procedures should attempt to evaluate if a student was confused about what he or she was supposed to do when performing in the testing situation.
- Students' motivation -- A test is not a valid measure of anything if the student is not motivated to perform in the testing situation. A test should attempt to assess independently how motivated the student was when he was performing in the testing situation.
- Students' misconceptions -- In order to help the student educationally, it is critical that tests diagnose in as fine-grain a manner as possible what inappropriate procedures or knowledge the student is using in performing any tasks. This kind of information is much more useful educationally than is knowledge about correct vs. incorrect performance or level of skill.
- Students' misinterpretations -- Often students may perform inadequately because they misinterpret what to do or misinterpret the language in a question. Tests should attempt to assess independently when the student has consistently made either kind of misinterpretation in a testing situation.

V. RESEARCH AND DEVELOPMENT

It is sometimes assumed, wrongly, that the temporal sequence is invariant: research comes first, and development comes later. That order sometimes occurs, but equally often one finds development coming first, or the two being inseparably intertwined. One had to develop space technology before doing research on moon rocks, one had to develop telescopes and radio telescopes before doing certain astronomical research, and building computers is an unending interchange between development and research. Recognition of the complexity of the relation is especially important in education, where, as Hawkins says, the wisdom of the best practitioners is frequently more advanced than the theoretical formulations of the best theorists. It is not always necessary, and not always desirable, to assume an invariable sequential order between research and development.

We have listed a number of research and development activities below. Central to our vision of a new learning and testing environment are the first two research and development projects listed below:

A. Development of Activities Integrating Testing and Teaching

Concurrently with research on student's processing and task demands, it is possible to begin developing activities and games for students that are educational. These should provide feedback to the student and teacher as to what the student is learning and what problems the student is having. Such an effort is being carried out for mathematics by Project Torque (Schwartz*). This should be extended to both reading and writing as well as other parts of the K-12 curriculum.

These activities and games can be either paper and pencil activities, as in Torque, or computer-based activities as in Plato (Davis) and some of the intelligent CAI work (Goldstein and Brown 1977). For example, in the reading area the computer game called Adventure involves a student in reading instructions, in order to explore a set of caves looking for treasure and fighting off enemies. Building a reading curriculum around such games would provide a highly motivating learning experience where the student's progress could be recorded in a non-interfering manner.

If it turns out that information-processing technology becomes widely available to schools--and some of us think this to be a real possibility--then a new approach to assessment becomes possible. Building upon such present work as the BUGGY program (Brown and Burton), one can achieve assessments not in terms of numbers, but rather in terms of a precise and detailed description of exactly how things are working, and exactly how and where a student is making errors. This is analogous to what one seeks from a physician, or from an auto mechanic. One does not wish to learn that one's health is 78%, or that one's auto has an age equivalent of 3.7 years. What is needed is an accurate and detailed description of what works and what doesn't work, and what form of intervention will improve things.

[Parenthetically, we recognize that this kind of "negative language" has sometimes been used within education to create a label for a non-existent disorder, and also has other abuses in its history--but the fact is that, where improvements are desired, one does have to be reasonably clear on what it is that one wishes to improve. Unfashionable or not, negative evaluations can be part of the road to something better--perhaps even an essential part.]

---

*Citations without dates refer to papers in Part Two of this volume.

If, then, computing power becomes readily available, the present concept of "testing" (for purposes of instruction) can disappear to be replaced by something quite different, that would exist on several levels. At the first level, could be a very extensive monitoring of student performance, resulting in a very large collection of data for each student. This might be called the "first level" of recording student performance, in the sense that relatively little selection, analysis, and transformation will have taken place. The "higher levels" of data would be obtained by selecting among "lower level" data, and by processing it in various ways.

As an example, the amount of review needed could be individualized for each student and for each topic. The selection and analysis of data for such a decision could easily be arranged.

Some "second level" data could be obtained from "first level" data, by following a pattern suggested by the BUGGY program that identifies any consistent error pattern in student responses to subtraction problems. This is possible in part because student errors very often show a consistent a consistent pattern. For example, one common error is to avoid "borrowing" or "carrying," by subtracting the smaller digit from the larger within the same column, without regard to the intended order of subtraction, as in

$$\begin{array}{r} 1928 \\ -\ 573 \\ \hline 1455 \end{array}$$

Clearly, it is not the answer that we care about, but rather the process by which that answer was produced. The student may never again encounter the specific number 1455 for the rest of his life, but if he continues to use his erroneous process for subtracting, he will get wrong answers more often than not. The Brown & Burton approach has two important virtues: (i) it shifts attention from "product" to "process," and (ii) it makes remediation more effective by showing the student exactly where his method fails, and where it needs to be changed. Brown & Burton achieve this by using a computer which holds, in its memory, nearly every wrong process for subtraction that students have been observed using. By matching student answers against the answers that would be produced by the various wrong processes stored in the computer's memory, the computer can usually identify exactly where the student is making his error.

Thus, whereas "first level" data consists of actual answers, "second level" data obtained by Brown & Burton's method is a specific list of things the students does incorrectly--not a listing of incorrect answers, but a listing of specific methodological errors.

B. Detailed Descriptions of Student Processing Strategies

Recently cognitive science has begun to produce detailed process accounts of how students carry out various tasks, such as those involved in mathematics (Brown and Burton 1978; Heller and Greeno), reading (Adams and Collins 1979; Bruce and Newman 1978; Rumelhart and Ortony 1977) and writing

(Bereiter 1979; Bruce et al 1978; Flower and Hayes, in press). With these same techniques it is also possible to pinpoint how students' processing can go wrong (Brown and Burton 1978; Collins and Haviland). These kinds of studies are necessary for constructing tests that look at process variables, and particularly those designed to diagnose suboptimal procedures students may use to accomplish any task.

We need to know what the mental representations are for the full range of topics we want to teach. In very limited parts of the curriculum, such analyses have already been done. (Greeno, on word problems; Brown and Burton on subtraction, etc.) The goal of this line of work is to create a precise model of the knowledge required to perform and understand a particular task.

Such models would include not just optimal or expert performance, but also procedures that generate novice performance (Larkin 1979); that contain partial knowledge (Klahr 1978), or "bugs" (Brown and Burton 1978). Thus the models could be used to diagnose student errors, and to determine not only that the student is wrong, but also what it is that the student (erroneously) thinks is the way to do something. The models would then be able to make instructional suggestions that make direct contact with the student's existing knowledge base.

At present we do not know what would be required to do this for even a single year in a single subject matter area (e.g., 3rd grade arithmetic). We also do not know enough about how to represent complex semantic relations that a "real" understanding of such subject matter would require.

Research is needed in the areas of detailed, fine-grain analysis of learner's performance (for both "model" and individual learners) in subject matter areas, on the underlying conceptual structure of such areas, on improved representational systems for describing and constructing complete knowledge bases, and on ways to assess and describe children's solution strategies (Klahr 1978; Klahr and Siegler 1978).

C. New Data Extraction Analysis and Aggregation Procedures

If we are to make use of the rich observational base proposed in our vision, we need new procedures for infering underlying cognitive structures from complex observations. Research is needed on how to improve methods currently confined to laboratory studies so that they can be used in the ongoing assessment of student knowledge. This, including that sort of error analysis already mentioned (e.g., BUGGY), would also extend to verbal protocols (Newell and Simon 1972), eye movements (Carpenter and Just 1977), reaction time (Chase 1978), move sequences (Baylor and Gascon 1974).

D. Studies of Children's Learning and Thinking Process

One research strategy in the study of mathematical and scientific topics is to focus on children's rudimentary forms of knowledge. These early representations are elaborated into increasingly sophisticated and mature forms through processes of learning and self-modification.

A promising area of research then, in the fields of both developmental psychology and artificial intelligence, is the study of self-modification processes (Langley, 1979; Anzai and Simon 1979; Sacerdoti 1977). More work is needed which focuses on the way that new information is acquired.

E. Computer Security and Privacy: What to Record, What To Keep

Many of the issues involved in avoiding undesirable consequences of a computer-based monitoring facilty are already receiving attention in the area of financial and military systems, large-scale university time-sharing systems, medical-record keeping, government auditing, etc.

"Research" in this area probably will first have to be on formulation of possible scenarios, followed by policy formulation, then followed by investigations of technological requirements to implement desirable policies with appropriate safeguards.

F. Research on Teacher's Understanding of Student's Knowledge

An extreme and ancient form of a test that teaches is the Socratic method in which teaching consists *entirely* of testing. Research by Collins (1977) has begun to reveal the information processing strategies used by "Socratic" tutors in their attempts to decide what questions to present to the learner. Further work along these lines would attempt to capture the strategies used by teachers who must form models of their students' knowledge.

G. Research on Testing Students in Groups

Traditionally, tests have focussed on the learning accomplished by an individual student responding to a teacher and to educational resources such as books, materials, or equipment. However, the role of fellow students in promoting or hindering learning is apparent: students often study together, they compete or they cooperate. They may distract or they may stimulate. It is reasonable, therefore, to enable a teacher to maximize the help he provides to the ongoing learning of his pupils, that we construct tests that explicitly measure the way in which an individual student performs within a group of learners. Research is needed to construct tests of this kind, to find ways of identifying and classifying the kinds of group learning patterns that occur, and to measure that way in which the individual student contributes and responds to such group learning experiences.

H. Factors and Biases in Individual Decision Making

Regardless of the sophistication of future computer based observational and diagnostic systems, it is highly likely that teachers will continue to play an important role in making judgments about children. Two lines of work in psychology seem important and relevant to increasing our understanding of how people make judgments about other people and about uncertain events. The area of "person perception" in social psychology has attempted to understand the extent to which and the conditions under which, we attribute the causes of events to things intrinsic to the person, or to things inherent in the situation. Such work has already been extended to important realistic settings (parole decision making), and it might be productively extended to classroom judgments.

A related line of work from cognitive psychology is concerned with the study of "judgmental biases". People appear to be consistently influenced by certain irrelevant or misleading features of decision making contexts. They appear to use unreliable heuristics such as "availability", "representativeness", and others in reaching their decision. Such cognitive biases are undoubtedly present in current decision making by teachers about students; further study of these problems might lead to more rational and correct evaluation procedures.

I. Test Validation

Achievement tests have traditionally held that the major evidence for validity lies in the process of development, particularly the content specifications. The argument is that external validity criteria that can improve on this do not exist. For example, looking at correlations with other tests attempting to measure the same thing is usually not done. Obviously this step is of minimal value unless one is willing to assess the other tests on, a priori, more definitive measures. Some construct validity evidence is often collected, e.g growth in relation to instruction, but it appears that the kinds of relevant evidence one can obtain are very limited. What is needed are external criteria. Some suggestions of ways to find these criteria are to be found in the papers by Koslin and Schwartz included in this volume. Systematic development of these and other possible criterion measures is needed. The existence of generally known criterion measures for achievement tests -- even if very awkward and expensive to use -- would be very valuable.

J. Research on Current Uses of Tests

A vision based on the need to remedy problems which exist only rarely is not likely to get acceptance. Therefore research, investigations of the following kind are needed:

1. How do teachers interpret achievement test scores? How frequently do they treat aptitude and achievement test scores the same way? As having the same meaning?

2. How good are teacher-made tests? Are they reasonably valid and reliable? How do teachers use them? Do they measure the objectives taught? In what ways are they better than standardized tests? In what ways worse?

   How biased are they typically? Do teachers believe the scores and act on them in preference to those from standardized tests?

3. What uses, if any, do teachers make of standardized achievement test scores? What uses, if any, do they make of scores from published criterion-referenced tests?

4. What uses do teachers make of the data they have available from student productions such as homework papers and the like? How often do they record these data systematically and treat them like test scores? Do they tend to record only their summative evaluations of these products?

5. The same questions as in 4, above, for observational data.

6. A development task preliminary to the envisioned future. Development manuals for interpreting student classroom work in relevant subjects, e.g. reading comprehension. Develop general schema based on cognitive science which will facilitate teacher use of non-test data such as student products and classroom observations of student work for diagnosis and instructional planning -- these would be crude initial steps toward the more adequate computer-based data structures in the vision. In this stage the teachers (and perhaps the students and parents) would be given a manual of directions, forms for recordkeeping and the like.

   Develop "hands-on" tests and/or simulations of "real life" application of school learning.

   A very large scale effort along these lines is practical, given funds. Tests of these sorts are relatively easy to devise but are expensive and have limited generality. Teachers would have to use many of them -- many computer-based simulations available would help -- to develop general conclusions about student competence. However, easy access to many such tests, feedback to teachers and students, and the relevance of effectiveness of instruction becomes possible. Only with substantial federal funds will such tests be created.

7. How do students typically react to standardized achievement tests? To teacher-made tests? How frequently do they feel threatened? Challenged? What conditions tend to produce these reactions? (Teacher attitudes, test administration procedures, etc.).

REFERENCES

Adams, M.J. and Collins, A. (1979) A schema-theoretic view of reading. In R.O. Freedle, ed., New Directions in Discourse Processing. Norwood, N.J.: Ablex.

Anzai, Y. and Simon, H.A. (1979) The theory of learning by doing. Psychological Review 86:124-140.

Baylor, G.W. and Gascon, J. (1974) An information processing theory of aspects of the development of weight seriation in children. Cognitive Psychology 6:1-40.

Bereiter, C. (1979) Development in writing. In L. Gregg and E. Steinberg, eds., Cognitive Processes in Writing. Hillsdale, N.J.: Lawrence Erlbaum Associates. (Also reproduced in the present volume)

Bruce, B., and Newman, D. (1978) Interacting plans. Cognitive Science 2:195-233.

Bruce, B., Collins, A., and Rubin, A.D. (1978) A cognitive science approach to writing. Technical Report No. 89, Bolt Beranek & Newman.

Carpenter, P.A. and Just, M.A. (1977) Reading comprehension as eyes see it. In M.A. Just & P.A. Carpenter, eds., Cognitive Processes in Comprehension. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Chase, W.G. (1978) Elementary information processes. In W.K. Estes ed., Handbook of Learning and Cognitive Processes. Hillsdale, N.J.: Lawrence Erlbaum Associates

Collins, A. (1977) Processes in acquiring knowledge. In R.C. Anderson, R.J. Spiro, and W.E. Montague eds., Schooling and the Acquisition of Knowledge. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Flower, L.S., and Hayes, J.R. (in press) Plans and the cognitive process of writing. In C. Frederiksen, M. Whiteman, and J. Dominic, eds., Writing: The Nature, Development, and Teaching of Written Communication. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Goldstein, I.P. and Brown, J.S. (1979) The computer as a personal assistant for learning. In J. Lochhead and J. Clement, eds., Cognitive Process Instruction. Philadelphia: Franklin Institute Press.

Klahr, D. (1978) Goal formation, planning, and learning by pre-school problem solvers, or: 'My socks are in the dryer.' In R.S. Siegler, ed., Children's Thinking: What Develops?. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Klahr, D., and Siegler, R.S. (1978) The representation of children's knowledge. In H. Reese, and L.P. Lipsett, eds., Advances in Child Development. New York: Academic Press.

Langley, P. (1979) A production system for the induction of mathematical functions. Behavioral Science 24: 121-139.

Larkin, J.H. (1979) Models of strategy for solving physics problems. Paper presented at the AERA Annual Meeting, San Francisco, April 1979.

Newell, A., and Simon, H.A. (1972) Human Problem Solving. Englewood Cliffs, N.J.: Prentice Hall.

Rumelhart, D.E., and Ortony, A. (1977) The representation of knowledge in memory. In R.C. Anderson, R.J. Spiro, and W.E. Montague, eds., Schooling and the Acquisition of Knowledge. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Sacerdoti, E. (1977) A Structure for Plans and Behavior. The Artificial Intelligence Series. New York: Elsevier North-Holland.

TEACHING, TESTING, AND LEARNING: THE APPROACH THROUGH RESEARCH ON CLASSROOM PROCESSES

Report of Committee 2

Sylvia Charp
Ann Cook
Lawrence Frase
Norman Frederiksen, Chair
Asa Hilliard
Judah Schwartz
Susan Stodolsky
Sheldon White

CONTENTS

I. INTRODUCTION 358

II. FRAMING IDEAS 359
- A. Informing Teaching and Learning 359
- B. A Collaborative Research Model 360
- C. Axioms and Guidelines 361

III. RESEARCH PROPOSALS 362
- A. Studies of Classroom Processes 362
  - 1. Introduction 362
  - 2. Illustrative Studies 363
  - 3. Investigations of Children's Learning Styles 364
- B. Educational Uses of Technology 365
  - 1. Introduction 365
  - 2. Educational Functions of Technology 366
  - 3. Research and Development Considerations 368
- C. Tests for Teaching 370
- D. Studies of Existing Tests and Test Use 371
  - 1. Tests and Their Effects 371
  - 2. Separation of Effects of Teaching from Maturation 373
  - 3. Test Use in Schools 373

I. INTRODUCTION

Tests as they are now used serve a number of different purposes. The committee focused its attention on one class of purposes, the use of testing to assist in classroom instruction and to further its ends. Some of the other purposes towards which tests are directed tend to make testing instruments at times irrelevant to, and at times disruptive of, a classroom environment. The committee sought for ways to reduce existing frictions among teaching, testing, and learning.

We saw as a kind of ultimate ideal the possibility of instructional environments in which testing was integrated into the learning process. The gathering of evidence about children's performance ("testing") would take place as part of the natural flow of the classroom, and the evidence would present itself to teachers and/or learners as easy and meaningful guidance towards the directions in which instruction or learning might proceed. Existing tests fall far short of this ideal, of course, and no one can imagine that we will quickly and easily approximate this ideal in the classroom. We sought to find ways, in the short term, to modify testing so as to make it less intrusive and more constructive for the classroom.

In the several sections of the report to follow, we attempt to spell out our guiding conceptions and then to offer proposals for research and development that might help us towards the longer-term and shorter-term objectives.

The first major section of the report offers some planning ideas intended to guide future research and development. The second major section sets forth specific research ideas. Within each of these two sections, the report deals first with a longer-term perspective and second with the shorter term.

Following this plan the organization guiding the subsequent sections may be envisaged as follows:

Framing Ideas

The longer-term perspective : Informing Teaching & Learning
A Collaborative Research Model
The shorter-term perspective: Axioms and Guidelines

Research Proposals

The longer-term perspective : Studies of Classroom Processes
Educational Uses of Technology
Tests for Teaching

The shorter-term perspective: Studies of Existing Tests and Test Use

II. FRAMING IDEAS

A. Informing Teaching and Learning

It has been argued that present-day testing programs are largely extraneous to everyday classroom teaching. Achievement tests are given infrequently and offer relatively crude estimates of individual children's performance, so that teachers can take little practical guidance from them.

In order to address the problem of how "testing" or something like it might help in the classroom, we posed a hypothetical situation. Imagine that present-day standardized testing did not exist, that not only tests but all memory of them were to disappear from American society. Would we seek to reinvent testing for the sake of the teacher in the classroom? Do teachers want information, or feedback of a kind that now does not appear in classrooms? Or, alternatively, are there existing kinds of information "latent" in classroom activities whose value for teaching might be realized by programs of classroom studies?

What kind of "testing," in any sense of the term, could teachers use?

1. They can use better observation of the learning of all children before them.

Observations of classroom processes have repeatedly shown that teachers do not, and probably cannot, give equal observation and equal attention to all children in a normal-sized classroom. Some children interact with teachers much more frequently than others. What this may mean, for two children in a class of 35, is that one child may effectively experience a "pupil-teacher ratio" of 1:5 and another a ratio of 1:60. Research might try to bring teachers information that would allow them to encounter students more effectively.

2. They can use better understanding of the variety of children with whom they must deal.

Part of the uneven distribution of teacher time and effort to children may have to do with uneven understanding of children. Children differ and so do teachers. Some children within a classroom are more "comprehensible" to a teacher than are others. The teacher can make more sense of the child's efforts and knows better how to deal with the child. The teacher's grasp of other children's performance is more vague and unsure. The teacher may see that the child is having academic problems, or acting up, but the teacher has difficulty making sense of the pattern of the child's activity. Research can provide characterizations of learning styles, of patterns of performance, in individual children that enable the teacher to make sense of a child's activities.

3. They can use learning systems that provide feedback and guidance directly to the student, thus facilitating the tasks of the teacher.

The ultimate purpose of providing information to the teacher is to help the teacher to assist the child in learning. Well-designed educational systems, offered in print, graphics, and/or technological media, can effectively assist the teacher by directly providing feedback and guidance to the student.

B. A Collaborative Research Model

Given the nature of research that we believe to be important, a consideration of the way in which the research would be conducted seemed essential. While we do not intend to propose one "right" way of working, we have devoted considerable discussion and thought to a model of collaborative research.

Periodically, there have been calls for increased communication and exchange between researchers (usually university-based) and teachers. We are all aware of some highly successful and some very unsatisfactory instances of such efforts. There are and will continue to be priorities and concerns which are not identical for these groups. However, we believe the improvement of classroom learning and teaching *and* the improvement of much educationally directed research can be nurtured by genuinely collaborative research. We have in mind situations in which researchers and teachers work together toward an understanding and improvement of classroom process and/or children's learning. Under the usual circumstance, both parties would spend time in the classroom and would define mutually important problems. Such collaboration should be mutually beneficial -- it should "ground" the work of researchers in the realities of classroom life and should help teachers better discern systematic patterns in children's learning and functioning.

There are a few current projects using collaborative approaches which we believe are good illustrations of what we have in mind. These include The Ford Teaching Project (John Elliot and others) and the ETS Documentation of Reading Project (Bussis, Amarel, and Chittenden).

A major supposition of ours is that teachers' perceptions of children, subject matter, and the process of teaching have substantial impact on their behavior in classrooms. Further, we believe that current standardized testing procedures often impinge in undesirable ways on the teachers' views and actions. We want to encourage the development of interpersonal and technological ways of increasing teachers' skills and vision and enhancing learning in children. We also want to foster research activities which will be better informed by the complexity and the real ecologies of classrooms. So the researcher's vision, too, must be stretched to encompass the real world of schooling.

C. Axioms and Guidelines

In thinking about the general problem of new approaches to assessment the committee made certain assumptions which may be thought of as axioms. These are:

1. There is an asymmetry of allowable inference; i.e., if a child performs some task it may be possible to infer that the child is competent with respect to that task. If the child does not perform the task, no inference can be made about the child's competence. Reports of test outcomes, therefore, should be of the form "The child can do ..." or "The child didn't do ..." but never "The child can't do ..."

2. People are different. Schools should make them more so. If there are certain critical thinking skills we want all children to develop we should allow the children to demonstrate their competence with these skills in subject areas that interest them.

3. The reason a person offers in support of an answer to a question or resolution of a problem is an important piece of information. Indeed it may not be possible to understand the response without an appreciation of the reason it was offered.

Further, the committee felt that assessment procedures and instruments for classroom purposes ought to be shaped with the following guidelines in mind:

1. Assessment procedures and instruments to be useful in the classroom must have diagnostic value. Moreover, the information they provide should be made available quickly to both students and teachers.

2. Assessment procedures and instruments should be flexible and able to accommodate the idiosyncratic learning styles of different children.

3. Cultural, ethnic, and linguistic assumptions in the design of procedures and instruments are inevitable. They ought to be explicitly stated. For this and other reasons tests should be openly available for examination by parents and others interested in the instructional process.

4. Any assessment necessarily probes a child's understanding of both the content and the language of the question. In devising new assessment procedures and instruments we undertake the responsibility of minimizing this veritable confounding.

5. Assessment procedures and instruments should be designed to be motivating, and to elicit the children's best efforts.

6. It ought to be recognized that particular assessment procedures and instruments may not be appropriate for particular groups of children; e.g. written instruments should not be used in the earliest grades.

There is one further guideline we wish to assert that differs somewhat in kind from those given above.

Minority groups have long suffered major negative consequences from studies of student behavior done by investigators unfamiliar with the cultural group being studied. There is abundant literature to show that cultural misunderstanding is a common source of error in assessment. In order to minimize the possibilities for continued misunderstandings, the criteria for funding research by NIE should include one that there be evidence of appropriate cultural familiarity and sensitivity.

III. RESEARCH PROPOSALS

A. Studies of Classroom Processes

1. Introduction

We discuss here assessment as an inherent part of the classroom process. We wish to focus on such assessment as provides information that helps teachers teach and helps children learn. We are not concerned here with assessment that is intended to inform placement and selection decisions, nor with tests or other assessment procedures meant to inform accountability decisions. On the other hand, we believe that the assessment methods we might evolve to inform the teaching and learning process will also be helpful for informing parents of their children's attainment, styles, and strengths.

In order to proceed to develop assessment procedures and instruments that are, in fact, useful to the teacher and student, we must evolve from the classroom itself a theoretical framework within which we can build descriptions that capture the richness and complexity of the learning of the children both individually and collectively. During the course of the school year an immense amount of information is accumulated in the minds and notebooks of all of those in a classroom. It seems to us that assessment procedures designed to help teaching and learning should be able to draw on this resource.

How do we move toward this goal? Clearly if we are successful in devising ways to use these data that come from the daily lives of the people in the classroom, we are going to have to pose and answer a variety of questions. These questions constitute a research agenda that can and should be addressed in the milieu in which its outcomes are intended to be useful. Doing research in the classroom will, we believe, lead inevitably to the blurring of the boundaries of such categories of activities as teaching, learning, and testing. Consider, for example, the exchange among a teacher and a group of children. Can one identify the parts of that exchange that constitute teaching and distinguish them from the parts that constitute testing? Or, for example, can one reasonably distinguish teaching from learning in the children's conversation among themselves? The very categories, teaching, learning, and testing, we assert, are not such as would grow out of the life of the classroom.

Classrooms are complex places. They are complex not only because they have collections of individuals who are each complex in themselves. They have a further complexity that differs for the complexity of the individual. It is crucial that any research addressed to devising better ways to look at the classroom and the people and the events within it respect such complexity.

Complex settings are difficult to describe. The temptation to aggregate complex data in order to convey economically the nature of the enterprise is great. Too often in the past, however, such aggregation has been simplistic and has lent itself to the perpetuation of social mischief and pedagogic malpractice. An important research question confronting us is the devising (from the data themselves) of ways of describing that are more economical and analytical than referring each time to the ever-growing body of raw data that inevitably keeps accumulating. By the same token, we need ways of describing that are more informative and insightful than percentiles or stanines. Ultimately, of course, powerful ways of describing will grow out of a much deeper understanding of the structure of the subject matter than we now have. But we need not wait. By and large, we do no use the knowledge that we now have to shape such descriptions. We should do so, and we should improve those descriptions as we continue to learn more about the ways in which children represent the content of the domains we ask them to learn.

2. Illustrative Studies

One important set of classroom process questions is essentially descriptive in nature. As we have indicated, descriptive materials are important starting points for much scientific work and for teaching. Descriptive studies can also be invaluable in providing ways of really knowing the course and character of American education over a period of years.

We reemphasize the complexity of classroom phenomena. We also need to convey the fact that productive, useful descriptions of classroom realities are difficult to obtain. Many views and viewers are often needed to arrive at an approximation of the reality. Theoretical insight is often central.

We recommend funding of studies such as the following:

- Collections of rich ecological descriptions and analyses of educational environments in a representative set of schools. Prior work by Roger Barker, Paul Gump, E. Goffman, Barbara Biber, and others could be used as a base for formulating data collection and analysis procedures.

- An analysis of setting characteristics and their relationship to the achievement of diverse educational objectives. It has been hypothesized (Grannis, Gump, Stodolsky) that different properties of school behavior settings may be optimal for different educational goals. Variables such as who is pacing the child's work, who chooses tasks, the types of materials used, the nature of feedback provided, and opportunities for social interaction seem to be important setting variables. Different combinations of these variables should be explored both naturalistically and quasi-experimentally in relation to the attainment of lower and higher mental process goals as well as affective processes.

- An investigation of classroom questions. One important framework for investigating classroom life and teaching quality could be studies of the questions both children and teachers ask in classrooms. Attitudes and conceptions of learning, normative rule systems, and role relations might be revealed through recording and analyzing social exchange among teachers and children and among children themselves.

- Children's behavior in schools of various kinds. Studies seem needed of how children select materials, workmates, and activities in different settings. How does the children's learning style affect his utilization of the learning environment?

- Children's use of various forms of feedback. We do not know enough about the ways in which children respond to feedback. Can children utilize detailed information about their learning and progress? How can they be helped to understand feedback sources?

- Longitudinal case studies of children's school learning in the early years. Case studies of how children arrive at levels of understanding by using differnt routes. These studies should include the development of reading, math, writing competencies, and affective processes.

- Analyses of classroom structure and children's styles. Longitudinal descriptive studies of children's work and styles of learning as they related to the learning and testing environments in which children function are needed.

- Impact of instructional means on long-range child outcomes. Studies of the ways in which means by which we have children learn subjects relate to the ends. Thus, does the learning of reading, say with a phonetic approach, affect the interests and abilities of the mature reader? Or does the use of pattern drill affect children's tendencies to be curious in other contexts? Are attitudes affected by the means used to teach?

3. Investigations of Children's Learning Styles

We know that different children learn in different ways. What are the varieties of ways children have of expressing knowledge? What is the nature of their preferred medium? How does such preference affect learning strategies?

a. Relations Between Children's Learning Styles and Preferences and Teachers' Behavior. These studies would explore the impact that children's styles and preferences have on the range of material made available by teachers and the variety of behaviors teachers use.

b. How Teachers Use the Concerns, Needs and Agendas of Students to Build Curriculum. The collaborative model discussed earlier suggests a vehicle for facilitating the feedback of initial research findings to improve instruction and, therefore, serves as a training function. A study of how teachers use the concerns, needs, and agendas of students to build curriculum and which was grounded in a triangulation procedure (involving students, teacher, and observer) might provide insights into specific ways curriculum can respond to the particular learning demands of students.

c. The Effects of Varying Teaching Styles and Types of Materials on Student Learning. Through an exploration of teaching styles and their intended and actual effects on students, teachers could gain skill in finding appropriate methods to extend student understanding. Such research would call for classroom case studies which rely on observation, interviews, and negotiated sharing of collected data among participants. (Here the work of Elliott, Stenhouse, Stake, McDonald, and Perrone would be helpful.)

d. What Kinds of Information About Children's Learning Affect Teachers' Teaching? We need to know what kinds of information about children's learning make what kinds of differences to teachers' teaching, over time.

e. How Does the Provision of New Ways of Thinking About Subject Matter Affect the Teacher's Teaching? Better understanding of subject matter is assumed to impact on teachers' instructional and diagnostic behaviors. Research is needed to explore in depth how and under what conditions these impacts are maximized.

f. What are the Relations Between Curriculum Content and Its Impact on the Provisioning Done by Teachers and Children's Learning Strategies? To what extent do assumptions embedded in particular curriculum materials determine the organization of the classroom, the selection of teaching style, and the response of students?

B. Educational Uses of Technology

1. Introduction

It is the purpose of this section to assume a broad view of technology and to consider technology as a tool which will assist students in the learning process. There is the further assumption made that all students can learn and if placed in the proper environment will learn. If a direct relationship should exist between learning, teaching, and testing, technology, in particular computer technology, does permit this relationship to be achieved. Evidence does exist that the effectiveness of this relationship needs to be examined so that changes, modifications, and improvements can be made. New developments in technologies and future directions in the use of technology must be planned for in order that the educational environment will best benefit from whatever changes will occur, and teaching, learning, and testing becomes part of a total process.

2. Educational Functions of Technology

Technology, broadly conceived as a variety of materials and procedures that can be marshalled in support of student learning, can facilitate and expand motivations to learn, information acquisition, presentation and data collection, and it can expand the use and development of intellectual skills. The motivational, informative, and skill support functions of technology are briefly described below. We do not conceive these to be mutually exclusive areas, for acquiring information can have motivational properties, and being motivated can result in the acquisition of skills. For purposes of simplication, we treat these three functions of technology separately.

a. Motivation

The motivational effect of technology has already been observed. The learner who has been placed in an environment which is individualized, responsive, and non-threatening, and who is provided with immediate "feedback," becomes totally involved for long periods of time. It has been seen that social interaction occurs as students work together in pairs or in groups to discuss a presentation; for example, a computer simulation. The ability of technology to extend teaching and learning beyond the classroom provides opportunities for parents and the community to become involved in the instructional process. The variety and richness of technological experiences that can be available have only begun to be explored. What is available today is still limited, but the future does hold out exciting new experiences for motivating the learner.

b. Information

A variety of materials and devices can be used to display information to students, teachers, parents, and administrators. How subject matter knowledge should be characterized, stored, and displayed are useful questions for research, but the capabilities of, for instance, new electronic technologies, go far beyond the mere visual presentation of static bodies of data.

The invention of economical large-storage devices, such as video-discs, makes possible access to large bodies of data. Entire libraries of printed materials can be rapidly accessed and probed for successively more detailed information. Children can explore at will the accumulated history of their culture and other cultures. This information comes to life when supplemented with animated images and sound.

Not every student reaches understanding in the same way. For some, visual images may hold special appeal and effectiveness. Not every student has the same background. The capacity to obtain data on student performance, to present information in different modes, and to modify the words used are some of the elements of adapting information presentation to students that technology can help with.

Ease of access and flexible displays can more effectively provide information to users. But the user, actively engaged in learning or casually exploring, can provide a wealth of data (e.g. on learning styles), if properly interpreted, to teachers, parents, administrators, and to the user as well, especially if these data represent a longitudinal and multidimensional perspective on human performance. These records provide descriptive data that can involve researchers intimately in the processes of classroom learning.

Educational programs and performance data can be transported easily across distance and time, providing information resources that have immediacy and historical potential.

c. Skills

The simplest use of educational materials is to provide drill and practice. Even this basic function exerts a forcing constraint on curriculum development; clarifying objectives, methods of instruction, and revealing design misconceptions.

But the potential of new electronic tools goes beyond drill, practice, and rote learning. New possibilities exist for supporting the skills that students need to assure progress in school. For instance, computational skills weakly learned restrain a child from acquiring more complex mathematical skills. But inexpensive devices can temporarily help students over computational hurdles, accelerating their progress. We do not understand enough about this kind of supporting role for new technology. New technologies may interfere, and we need to understand the dangers involved. The attraction of television perhaps has shifted our children's attention away from the exercise of basic skills. Properly coupled with detailed analysis of performance in problem-solving, the data collection made possible by continuous monitoring of student performance can be used to provide detailed training to teachers in the diagnosis of student learning processes. These diagnostic assessments, provided directly to students, can give to students control over their own learning processes.

The most intriguing aspect of advanced hardware technologies, for instance computers coupled with large-scale storage devices, lies in their potential for extending the reach of human intelligence. The implications here are only dimly seen, but if our assessment is correct we stand at the threshhold of a revolution in productive human thought. New technologies, large scale and small (personal computers), can be used to produce complex functions that test mathematical hypotheses (too difficult to calculate by hand), to display related information simultaneously (facilitating synthesis and reasoning), to produce musical or visual scores only outlined and structured by the human user. In this regard, new technologies serve as tools that eliminate the repetitiveness of intellectual work, well mastered, that is necessary in the production of complex literate and scientific products.

3. Research and Development Considerations

a. Process

It would be advantageous to build upon information that has already been acquired in the area of educational uses of technology. Activities using various forms of technology are going on both nationally and internationally. Whatever has already been learned should be shared and new areas explored. A multi-disciplinary approach, that is, collaboration among researchers, psychologists, technologists, curriculum developers, teachers, and so on, must be initiated and maintained over long periods. This type of activity requires long-term support. Short-term school projects do not provide sufficient evidence about effects on learning. Longitudinal studies are necessary. Alternative plans for continuous effort on the use of technology need to be explored and funded. This involves educating not only the educational community, but also parents and the community at large, on the advantages and proper uses of technology. It would be advantageous for a National Center to be established which would investigate and be concerned with the following areas: Basic Research, Applied Research, Development of Models, and Dissemination Procedures.

b. Illustrative Projects

(1) Motivation

- Conduct research to determine the effects of different kinds of feedback on student performance and motivation.
- Conduct research to determine the ways in which subject matter can be adapted to different student interests.
- Conduct research to explore the ways in which social exchanges and cooperation via computer can be used to motivate learning.
- Conduct research to explore the ways in which student attitudes toward the use of instructional materials and devices vary with duration of exposure and whether such attitudes transfer to other educational contexts.
- Investigate the features of instructional environments (e.g. visual displays) that affect student interest and the maintenance of attention.

(2) Information

- Develop prototypic systems that allow students to interrogate data banks for specific subject matter information at varying levels of detail.

- Conduct research on the use of small computers to keep detailed personal diaries of student problem-solving skills and other activities over long periods of time.
- Investigate the effectiveness of different technological products (e.g. videodiscs) for different instructional purposes.
- Conduct research on the optimal range and duration of information on students that is needed for instructional purposes.
- Investigate the ways in which technology can be used to adapt to different learning styles.

(3) Skills

- Develop detailed methods for describing and diagnosing successful and unsuccessful reading and mathematics performance.
- Investigate the kinds of skills that teachers need to use technology effectively.
- Conduct research on the role and effects of technology in extending students' abilities to generate complex products (e.g. mathematical functions, musical score, text).
- Investigate the role of technology to support specific performance in areas in which a student has a physical or problem-solving deficiency.
- Explore the role of technology as a management tool in making decisions about class size, assignment of students, and so on.

3. Outcomes and Conclusions

We have tried to provide a new perspective on the educational uses of technology. Basic to this conception are two assumptions: (a) that the beneficiary of data collection in the school should be the student; and (b) the instructional context can provide boundaries for basic research questions that will ultimately advance and bring together research and practice. We believe that basic research will flourish under the tutelege of real classroom problems and data; and classroom instruction, and consequently our students, will also flourish. Technology is the catalyst that encourages this transformation. In order to achieve the goals of this program, a variety of disciplines must work together.

C. Tests for Teaching

Research in the following areas is needed to improve the use of tests in teaching.

1. Effects of Different Kinds of Feedback to Students

Feedback to students can take several forms: (1) information that a response is right or wrong; (2) a critique of the response (e.g. "You made an error in reading the table."); (3) presentation of one or more models of good responses; or (4) a combination of the above. Experimental investigations of the effects of various kinds of feedback would be useful in guiding the development of tests for teaching.

2. Influence of Test Formats on Learning

Certain test formats may be preferred in certain situations; for example, use of a computerized test might be preferred to paper-pencil format in schools where the equipment is available. What difference does it make to a learner whether he watches a CRT and presses buttons or scans pieces of paper and makes responses with a pencil? Investigations of the effect of test format on learning and on possible side effects should be investigated.

3. Validation of Teaching Tests

Tests for teaching should be validated, both as tests and as teaching devices. As tests, there is need for studies of content and construct validity just as for conventional tests. In addition, they should be validated from the point of view of teaching effectiveness. Methods for carrying out such studies should be developed.

4. Process Models for Teaching Tests

Studies aimed at developing and testing models of the cognitive processes involved in taking a teaching test should be carried out. Such models would be particularly useful if the component processes were interpretable in terms of remedial actions.

5. Individual Differences in Learning from Teaching Tests

Children are likely to differ with regard to the patterns of behavior they display, or the processes involved, in learning from a teaching test. For example, some children may spend a relatively large amount of time in preparation and planning and a small amount in random or trial-and-error behavior. Since remedial methods may depend on such differences, information regarding different patterns of learning would be useful to the teacher. Studies should be carried out aimed at dicovering such individual differences, their significance in learning and teaching, their correlates (e.g. personality characteristics), and their generality.

D. Studies of Existing Tests and Test Use

It is recognized that a program of research of development such as we have proposed will take years to complete and that a system involving tests for teaching will not be in place for many years. It is therefore likely that tests of the sort now available will continue to be used. The suggested projects that follow are intended to provide for continued examination of conventional tests in order to understand better their characteristics and their influence on education, to enable us to draw from them what is useful, and in order to suggest modifications that make them more useful for such purposes as accountability and system evaluation.

1. Tests and Their Effects

a. Studies of the Content Validity of Selected Standardized Tests of Educational Achievement

Content validity has to do with the question of how accurately an achievement test reflects the content of the domain it is supposed to represent. It is suggested that content domains be compared using three definitions of the domain: (1) the results of a systematic survey of classroom teaching, textbooks, syllabi, and course outlines used in teaching the domain; (2) the opinions of a panel of highly competent experts in the field (including teachers) as to what the domain *should* include; and (3) the items in the test.

b. Studies of the Predictive Validity of Selected Standardized Tests.

The usual criterion used in studying the predictive validity of an educational test is subsequent academic performance as measured by course grades. Most users of tests, however, probably assume that the tests are valid for other purposes, such as predicting success in careers relevant to the area of the test. It is proposed that criteria be constructed that reflect relevant aspects of career success and that studies be conducted of test validity using such criteria as well as conventional criteria. An example of relevant criteria for examinees who become teachers might be measures that reflect gains by children in courses taught by the examinees.

c. Comparison of Test Scores With Measures of Social Background (e.g. socioeconomic status) as Predictors of Academic Performance.

Test performance is substantially correlated with items of information that reflect the social or economic status (SES) of the child's parents. It is sometimes alleged that measures of SES are as good as test scores in predicting educational achievement and that test scores for prediction are therefore unnecessary. Studies are suggested in which relationships of various indices of SES to school achievement are compared with relationships for test scores.

d. Studies of the Construct Validity of Selected Standardized Tests.

Construct validity has to do with the extent to which scores on a test respond to experimental treatments and are related to other variables in ways that are consistent with a theoretical conceptualization of the test. The other variables may include other test variables, criterion variables, personality and background variables, situational variables, and the like. Thus the construct validation would include proposed studies 2. and 3., above. Variables should also be included that represent other constructs that the test in question should not measure, according to the theoretical conception of the test. It is suggested that studies of the construct validity of selected standardized tests be carried out.

e. Devise a Method for Discovering Whether the Existence of A Standardized Testing Program in a School Makes Any Difference in What Goes on in That School.

Critics of tests often assert that testing program have no effect, that score reports are merely filed and forgotten. The idea is to devise a method or methods to discover whether or not the existence of a testing program makes any discernible difference in any educational outcome. Design problems for such an investigation are obviously difficult, as are problems of choice of criterion data. In view of the many factors in school, home, and community that influence school achievement, it is unlikely that the effects of testing could be detected by such gross measures as mean scores on achievement test. Efforts to identify more subtle effects, possibly those concerning subgroups of children, may be required.

f. Devise Methods for Training Teachers in Making Better Classroom Tests.

Most teacher-made tests are either amateurish "objective" tests (true-false or multiple choice) or essay tests. The objective tests are likely to stress knowledge of detailed facts, while the essay tests tend to permit the student to write whatever he knows that is somehow relevant to the question, and they are therefore very difficult to grade fairly. The middle ground that could be taught is how to prepare examination tasks that impose such constraints on the examinee that all are performing the same task, and how to develop scoring methods that yield more reliable scores and that reflect not just overall quality but all important aspects of performance.

g. Develop Methods that Permit Use of Constructed Responses in Large-Scale Testing Programs.

It is not possible to elicit performance that reflects originality, ideational fluency, or creativity using any of the usual machine-scorable formats. It is therefore important to devise methods that make possible the use of free-response formats in large-scale testing. One possibility is two-stage testing, in which the examinee writes his response in stage 1 and in stage 2 reports what was done in machine-readable form (with suitable precautions against cheating). Another possibility is to use a computer to recognize key words that suggest the nature of the idea written by the candidate.

h. Use Matrix Sampling Designs for Obtaining Information Regarding Performance of Groups of Students.

Information about performance of groups of students (schools, classes, regions) is useful for accountability, administrative purposes, and informing the public about the state of the schools. The information is typically obtained by giving tests to all the pupils and reporting such statistics as means and dispersions of scores. The time and effort would be greatly reduced by sampling both pupils and test items, so that many small random samples of students take subtests composed of samples of the items. A consequence of the economy of this procedure is that it is not necessary to use machine-scorable tests. If appropriately designed tests that require constructed responses were used, the evaluations of schools could be based on data that are free of the bias against creativity that characterizes machine-scorable tests.

2. Separation of Effects of Teaching From Maturation

There is increasing knowledge about how humans think (cognition) and how humans function in social groups. In many cases, the new knowledge is important potentially for instruction. However, it is also the case that the knowledge of how a learner functions does not always require a response or an intervention from teachers. Sometimes educators try to make maturational thinking processes (which develop automatically) a goal of instruction. They even design school experiences so that these maturational skills can be practiced. Presumably the goal is to make them develop earlier or "better." It becomes necessary to make a distinction between functions which reflect maturation, which occurs in a supportive environment, and those which develop in response to direct teaching. Behavior which reflects maturation may require a response in instructional design which harmonizes with but does not attempt to change a learner's developmental progress. For example, it would be wrong to try to teach the Piagetian function of "object permanence" if it will occur all by itself. It is possible, with difficulty, to teach a concept to a young child that can be taught easily to a somewhat older child. An important function of testing might be to determine when a given child has reached the maturational level that makes the learning relatively easy for the child. Research is necessary to distinguish as clearly as possible those places where teaching intervention may be suggested by developmental knowledge.

3. Test Use in Schools

a. Design Paradigm for Ecological Description of a School

The school setting exists in an "ecosystem." The meaning and utility of data are related to their location within the ecosystem. Testing and evaluation, therefore, take on added meaning and validity when considered within a whole system. Research is necessary to develop models of data gathering which permit systematic and simultaneous consideration of important dimensions of an educational ecosystem. The failure to collect certain situational information along with "testing" will cause later attempts at interpretation to be confounded due to the loss of essential data.

b. Study of Test Use by School Personnel

A more systematic and sophisticated description of the actual pattern of test use by teachers, school support personnel, administrators, and others is needed. A simple survey would be inadequate. Some form of systematic observation is essential. In the absence of this information, it is impossible to determine actually how useful tests are. A survey of literature on test use, and a new indepth observational study is needed.

c. Study of What Test Users Know about Testing

It is important to know what the primary users of existing standardized tests know about the properties of those tests or about the general nature of standardized testing. For example, it is impossible to determine whether the influence of tests on instruction is due to the tests themselves, to the use of tests, or to neither. A review of literature on user knowledge of "necessary information" about tests should be done. Further, actual tests should be given to a sample of users to determine the general level of understanding to be expected.

d. Study of Teacher Training in Test Use

The source of user knowledge about standardized tests needs to be examined. The improper use of standardized tests may be due in part to the lack of user preparation. This lack of preparation may be due to the absence of content on standardized tests in teacher education curriculum. A content analysis of present teacher education curriculum is needed to determine what training on standardized tests is included.

e. Identification of Models of Exemplary Test Use

It is possible that exemplary use of standardized tests can be found. If models of exemplary use can be found, they should be described in detail. A study is needed to locate and to describe sites where the proper use of standardized tests can be shown.

HARD DECISIONS FROM SOFT DATA

Report of Committee 3

Alonzo A. Crim
Parker Damon
Barbara Hayes-Roth
C. Dalton Jones, Chair
Bertram L. Koslin
Andrew C. Porter
Frances Quinto
Maria Ramirez
Ross Taylor
Jerrold R. Zacharias
Ralph W. Tyler,
ex officio

CONTENTS

I. PREAMBLE 376
7 Steps for the NIE to take.

II. NATURE OF DECISIONS AND DATA 378
A discussion of one purpose for collecting data, lists of other purposes and sources of data, and an example of the limited scope of information.

III. DECISION-MAKING CONTEXTS 380
A description of how the Atlanta Public Schools use test data, and the test related issues practitioners encounter.

IV. RESEARCH ISSUES 387
Recommendations for the NIE to promote research in two areas, and specific questions which need answers.

V. CONCLUSION 389
Where the NIE should begin.

I. PREAMBLE

> "Ask me no questions and you'll be smart.
> -Kelly Bice, grade 3, Good Hope, Alabama[1]

Anyone who wants to improve tests and their use must have guts. Our subcommittee believes unequivocally that it is not enough to say that tests have weaknesses and are misused and then go on to other matters. Someone has to say, "Stop'" "Look at what is going on!" "Make some changes." We charge the NIE with this task. We believe the people at NIE have the courage to do so. We believe the NIE can assume major responsibility and play a decisive role in initiating a new era in educational assessment.

Specifically, the NIE should do the following:

1. Make explicit to everyone (pupils, parents, public and professionals of all kinds) that a person's abilities, activities, and attitudes can not be measured. The public, especially, misperceives that hard data exist, and that test scores constitute these data. The public does not realize how quickly the point is reached where we do not know how to discriminate validly among people, but where the data mislead us to think we do. This is what is meant by the myth of measurability."[2]

2. Explain to everyone the inherent dangers involved in the use of tests: from doing what they cannot to saying what they should not (i.e., labelling). Within the context of decisions to allocate scarce resources, there comes a point beyond which we no longer know how to discriminate validly among applicants. In such cases, test data should not be used as an easy way to make decisions beyond their validity. When the NIE supports the use of tests, it should require that warnings be printed on the tests such as, "Danger, the contents of this test may cause you harm."

3. Develop parental and public understanding of tests through open and frequent discussions of the quality of individual tests and of how different tests compare. Therefore, the NIE should make clear that it will not submit to test secrecy - by advocating a policy of open access to tests.

4. Describe how the use of tests fosters institutional racism, stop these practices itself, and advocate that other institutions stop also. The NIE must expose the way in which society as a whole condones the deciding of difficult issues on the basis of faulty information. This requires a frank, public analysis of how decisions are made and how some members of the population, especially the minorities, are disportionately affected by these decisions.

---

[1] As reported in the Boston Globe, March 10, 1978.

[2] Houts, Paul L. (Ed.). The Myth of Measurability. New York: Hart Publishing Co. Inc., 1977.

5. Discuss the development of educational standards and the criteria for determining whether or not they have been met. Implicit in this discussion should be an analysis of how values conflict and contradict as well as change and support each other. What are important values to the test maker may not be for students and the many users of test results.

6. Advocate diversity in test construction; the content of tests should encourage expression of an individual's uniqueness and ability by asking the "What do You Know?" type question in addition to the "Do You Know What I know?" type.[1] The NIE should describe to the public how this can be done. In addition, the NIE should set forth the qualities or characteristics assessment and evaluation instruments and procedures should have.

7. Promote the use of tests which permit the test-taker to make a wide range of choices in test formats, and argue for the discontinuance of tests which by their requirements cause regimentation. Embedded within this issue are several others such as the need for continuous inservice support for professionals to make these choices possible, and the need for a confrontation of the many legal issues that are involved, particularly those related to open access, record keeping, and prior consent. The NIE should initiate this confrontation.

These are some of the concerns the NIE must bring before the public as well as lead professionals to tackle. The committee realizes that the seven issues overlap and represent only some of the host of vital questions and problems all educators should be aware of and concerned with. They are the links of the chain connecting buoy to anchor. Unattended to, the chain will rust, the weight of research will be lost, the channel will become unmarked, and vessels will founder.

These seven concerns are embedded in, surround, and join the more specific points discussed and suggested throughout this report. Separately, they may be likened to the blasts of a fog horn warning of danger, causing a response. Together with the other parts of the report they should serve as a beacon.

This Preamble outlines seven of the most important issues involved with the complexities of evaluation, assessment, and testing, and charges the NIE with the task of keeping these issues in the forefront of all its activities.

In the second section we discuss some of the types of hard decisions one encounters and the sources of data appropriate to them. This section, which focuses primarily on the issue of Public Reporting, stresses the interdependence of assessment data and the purposes for which they are used.

---

[1] See J.R. Zacharias' paper (page 58) which describes how a test may be one third mandatory, one third permissive, and one third a little of both.

The third section describes the influences of test data in Atlanta, Georgia, and the many issues teachers and principals encounter when using various forms of assessment information.

The concluding section strongly recommends where NIE's research endeavor should begin.

II. NATURE OF DECISIONS AND DATA

The many purposes for collection data, and the various ways of doing so are interrelated and form a complex system. Lists A and B give a sampling of purposes and representation of alternative sources of data.

List A: Purposes of the Data

1. Deciding what to teach - curriculum decisions
2. Deciding competence criteria: (including the economic viability of competencies decided upon)
3. Diagnosis (what is working, what is not, and why.)
4. Evaluation (of people, programs, institutions, systems, etc.)
5. Accountability (quality control? related to 4 and 8 below)
6. "Gatekeeping" - sorting, placement, selection
    - -credentials
    - -licensing
    - -entrance requirements
7. Allocation of resources
8. Public reporting
9. Job placement/opportunity

List B: Sources of Data for Making Hard Decisions

1. Teacher-given tests and exercises
2. Standardized tests (aptitude and achievement) - see process statements below for why this may be necessary as a short term source.
3. Individual student tests (vs. group test in 2)

4. Student questionnaires

5. Competency tests

6. Teacher recommendations

7. Systematic observation

8. Grades

9. Interviews

10. Products/performances

11. Criterion-referenced tests

12. Student histories (schools attended, communities, S.E.S.)

13. Selected recall of casual observation (almost anecdotes and gossip but more seriously considered)

14. Need sources of information from economic sector about competencies or performance criteria.

There are more, and this list is not ordered. While it should be cautioned that more is not always better, multiple sources of selected and relevant information should be available to and considered in decision-making processes.

The nature of the complex system is illustrated by Chart 1. One area of hard decisions, that of Public Reporting, shows how the public may acquire information, the nature of the information, the forums in which the information might be used and, finally, raises research questions indicating where alternative forms and sources of information may be acquired. Neither the lists nor the chart are thorough or exhaustive. The chart suggests that there are several critical levels or settings which presently exist and which should be considered in attempts to provide and gain information (research) pertaining to both the kinds of hard decisions and the data used as a basis of such decisions.

Example: Public Information at the National Level

Public reporting extends to the American school enterprise as a whole and is addressed to American society as a whole. In our list of purposes of data, public reporting is often related to such other purposes as evaluation in its various manifestations, diagnosis, and allocation of resources. Regarding our list of sources of data, List B, when it comes to public reporting and related purposes, the source of data is limited largely to standardized tests, when it should be comprised of much more meaningful data such as systematic observations, interviews, products/performances, and student histories. How do we remove the false dignity attributed to standardized scores and promote the legitimate informational value of these other data sources.

The nationally standardized tests are given at various levels in the elementary and secondary schools. They include, for example, Iowa Test of Basic Skills, the Stanford Achievement Test, the California Achievement Tests, and the Scholastic Aptitude Test of the College Board. College entrance tests may not ordinarily be thought of as school measures, but student scores are aggregated by school, as school measures, and nationally as a national indicator. Indeed, it is the decline over the past decade or so in SAT scores that has shaped most of the national debate concerning the quality of education.

In seeking to gain a national picture, to understand such trends as the decline in SAT scores and its causes, researchers do compare the results from the different batteries of tests. What we wish to suggest here as a research task for NIE and other government and private agencies, is to find out how to use for the same purpose additional measures: systematic observations, interviews, products/performances, student histories. The task is to learn how both to legitimize, and to bring together to form a larger picture, actual descriptions of classrooms, schools, and school systems. Is instruction by rote or by inquiry method? Do students learn word lists or do they read books. If they read books, which books do they enjoy? Can teachers afford to care or are they mere functionaries in a bureaucracy? In short, what experiences are students, teachers, parents, and school administrators having in schools?

At present the hard decisions about schools are based on measures that are limited in scope and unchecked by independent means. Systematic observations and other new data sources will be useful in their own right in helping to understand and improve schools. They will also serve as checks on the conclusions reached through the use of the more familiar standardized tests, especially the idea that achievement measures provide a total picture.

III. DECISION-MAKING CONTEXTS

"Teach me the whole Torah while I stand on one leg." Hillel replied: "Whatever is hateful to you, do not do to your fellow-man. Now go and study."

Rabbi Hillel
Pirke Avot 2:5

One report card for public schools is standardized tests. Citizens tend to believe in test scores. School systems work hard to improve test scores and run the danger of allowing the tests to drive and constrain their systems.

To illustrate how important achievement tests are, a brief description of how they are used in one conference member's school system will be given here.

CHART 1: PUBLIC REPORTING

How the kinds of Decisions (purposes of the data) and Sources of Data Relate in the Case of Public Reporting

Existing Public Forums	What goes on	Existing Data Sources	Research Needs: Alternative Data sources.
Student/Parent-Teacher/ Counselor Conferences	Grades/performance discussed: - raw scores - curved letter grades - percentiles - grade equivalents. Social/motivational and personality characteristics assessed. Home practices recommended.	Teacher generated tests Text or Unit Quizzes Standardized achievement tests Selected recall of casual observation; student histories.	-What do these data mean to teachers? How do they use them to explain instructional decisions and prescriptions? -What do parents and students come away understanding? Are there better bases for these communications and decisions. -How is the curriculum seen as integrating into a whole coherent experience?
School nights/orientation Home-school association/ P.T.A. meeting	Facilities displayed. Curriculum described (subject, content, teaching method).	-Tests are maybe mentioned and exam schedule revealed--but not the tests themselves -General statements about expected learning (Competency criteria implied but not made explicit). -Curriculum committee subject matter booklets available, expressing objectives.	-How do perceived social/personality factors (race, S.E.S., style, etc.) influence performance outcomes and prescription? -How diverse a set of alternatives for choice exist? How many are desirable? -What do parents, students, teachers bring to these settings to help understand what is being said, question it, or evaluate it?
Home	School newsletters, memos, newspapers, media (TV and radio especially)	Content usually not described in systematic way except publication of annual reports of school and district performance on standardized tests and school lunch menus. tests, and school lundh menus.	-Social class, and ethnicity influences on attendance, what is and what isn't said? -What additional public reporting is needed? What does the community/parents/teachers/ school administration want to know? How can they get it? What new mechanisms are needed for communicating and informing?
School District School Board meetings	Budgets, policy, and political debates of: - expenditures - discipline problems - quality of education presented in non-systematic order.	-How much is spent -Grades on achievement tests -Anecdotal reports of problems	

Prior to 1975, test scores in the Atlanta Public Schools (APS) had never been publicly reported. Despite no report of scores, the school system was perceived by many citizens as a low achieving school system, primarily because of the dramatic racial change in the student population from 80% white in the late 1960's and 80+% black by 1973. The low standardized achievement scores in 1975 confirmed the opinion of many black and white citizens that the public schools were bad.

During the 1974-75 school year, school staff conducted town meetings throughout Atlanta in order to hear the public views of the schools and to find out what the public felt should be priority school concerns. 10,000+ citizens attended those town meetings, and indicated that the public schools should establish four priorities:

1. Improve significantly the reading, writing, speaking, listening and ciphering skills of students.
2. Assist students in handling their futures - job placement.
3. Allocate the system's resources to students equitably.
4. Improve communication between school staff and the public, between school staff and students, and among school staff members.

1975-76 was a start-up year. Each school staff developed a local plan of improvement. The plans were compiled in the four school areas and at the central office. Much hard work of staff in 1975 resulted in no systemwide gains in achievement scores, but there was no systemwide loss. What the 1975-76 testing program in all grades did provide was data on all students. School system research staff used the data to report to students, parents and staff on successes and failures of individual students, common errors of students by classrooms and local school profiles.

A new feature to the planning activity was added in the 1976-77 school year. A review committee composed of the system's finest reading specialists, regardless of work assignment, was organized in each of the four areas. These review committees advised local school staffs on their plans prior to the school staff submitting their plans to Area Superintendents for approval.

The Review Committees assisted local school staffs in examining their plans against test data and instructional skills given in the curriculum guides. Classroom plans were checked for appropriateness with classroom profiles, staff development plans, and requests were checked with classroom and school profiles.

Review committees met with local school staffs every three months to review progress and suggest modifications to the plans if needed.

Test scores improved systemwide by 10% in 1976-77. A similar improvement was obtained in 1977-78.

Parents and students were informed of test data through information sessions. In 1977-78, 68% of parents came to school for those sessions. Reports were mailed to other parents. Presentations on tests were made on several public broadcasting programs, WETV Channel 30. Paid newspaper advertisements were used to inform citizens on progress.

Over the several years that the program has been in effect, media have given the improvement greater coverage. Their reports and the communication efforts of local schools have helped to attract the enrollment of 1,000 private school students to the APS. One elementary school principal of a school located in a community with a high private school student population posted a sign near the roadway in front of the school reading: "Garden Hills Test Scores up 20%. If your child attended school here s/he would be here now."

This positive experience of the APS is tenuous. APS cannot continually expect a 10% annual gain in achievement scores. If APS staff does not adequately educate and inform its public on tests it stands the risk of losing public support and perhaps suffering the displeasure of the public in dramatic ways.

It is evident that staff, students and parents give great attention to the tests. Students are taught what they don't know and what teachers anticipate students will be asked on the tests. The quality of instruction is attendant to how good the items on the Iowa Basic Skills Tests and other tests are.

APS have earned more scholarships since 1975 ($1,900,000 in 1975 and $3,700,000 in 1978). A correlation can be established with greater emphasis on instructional skills. A similar correlation cannot be established between improved achievement as measured by tests and job placement of students. Placement percentages have improved only through petitioning and calling on employers. Additional job slots do not automatically follow successful placement. Merit seems to go for naught, therefore, encouragement of students for hard work in school does not come through good job placements for successful students.

The APS experience has been good during the past two years. But it has taken years, more than two, to move forward with regularity. It may take a decade with the current momemtum to reach the mythical national norm. This progress may be hampered or reversed if there is a major change in teaching-learning conditions, i.e., change of tests, financial support, school staff commitment and public tolerance-support.

Staff needs and must have stable conditions to bring about significant improvement in skills achievement of students. You don't realize gains when major social programs are being implemented: 1973-74 and 1974-75 court-mandated desegregation program in Atlanta. Instructional improvement must pervade the educational atmosphere for substantial gains to be obtained.

> And it came to pass, when David had made an end to speaking these words unto Saul, that Saul said: "Is this thy voice, my son David?" And Saul lifted up his voice and wept. And he said to David: "Thou art more righteous than I, for thou hast rendered unto me good, whereas I have rendered unto thee evil."
>
> 1 Samuel 24:17-18

Issues and Concerns for the Practitioner

If the practitioner is to make the best uses of the data sources in serving the needs of students, then pre-service and in-service education are necessary. Training should include, but not be limited to:

- Generating data
- Interpretation of data
- Reporting of data to students, parents, administration, school boards, and the public at large
- Making decisions based on data - decisions about student needs and schooling programs.

Presently, the practitioner, both teacher and principal, must administer externally imposed tests which generate a single numerical score and a few statistical data. These test data, at best, provide limited information about a student or about the success/failure of a program. The data are now inappropriately used to make some hard decisions about students and school.

There is a danger when governing bodies use the information from a single source to allocate funds to schools or for programs. This practice has the potential for causing considerable confusion and duplicity in a school.

If programs are funded in schools with low scores, would there not be a temptation to encourage absences of high achieving students on days when tests are administered? The reverse may be true when funds are again allocated for "gains made" based on school scores, or when future funds are to be eliminated on the basis of student gains. Results of a test can terminate programs and spell an end to funding for established teaching positions and helpful programs. A variety of indices should be used in making the hard decision of continuing or terminating funding.

Anxiety-Producing Aspects

The anxiety produced by the test taking situation and the results of a test were stated, almost too dramatically, in a CBS interview with a student who recently had failed a high-school competency test. There was danger of not being graduated (next year) with a regular diploma. When asked what would happen if she failed the test again, the student answered, "I'll kill myself."

Does this story indicate that testing procedures should be various and sufficient enough to provide both the practitioner and the student a way to uncover strengths as well as weaknesses and failures? It is important that practitioners be informed about all procedures and placed in a decision-making position to select the ones that will help students understand their own strengths and potential areas of development. What are the best ways for the practitioner to promote an improved self-image for those students who need it the most? Can this be accomplished by using informal or observational techniques?

Issue of Teacher-Evaluation

Recently, the legislature in Florida considered a bill, "which provided that, for teachers, failure to teach students to reach a minimum performance on basic skills is grounds for dismissal, suspension...and failure to identify teaching deficiencies is cause for action against a principal or supervisor".[1]

The New Jersey legislature recently determined that teacher evaluation is to include application of "indicators of student progress."

No single indicator and no single measure can be used to assess the effectiveness of a teacher with students. In addition to the need for multiple considerations in evaluating teachers, no schools, no practitioner should guarantee that school attendance will result in learning. Influences outside of school that determine a student's motivation for accepting the school's program cannot be ignored.

Whether or not legislatures pass laws connecting student progress to teacher evaluation, public attitudes are formed by publication of recent test results. Like a thermometer - as scores go up or down - appearances are that something in the schools is awry and finger pointing is the result. Placing blame directly on the practitioner does not promote trust and security.

If there is the possibility that teachers will be evaluated on the basis of a student's test(s), then teachers would indeed be foolish not to "teach to the test" and provide for each student, as much assistance as possible to ensure the best score(s).

Basic Skills

The public is now demanding evidence of greater achievement in "basic skills" and an end to "social promotions." With such demands and amid much criticism, teachers are left with no alternative but to ensure that the evidence of achievement is indeed provided.

---

[1] Evaluation Panel (Ralph Tyler, Chairperson), The Florida Accountability Program: An Evaluation of Its Educational Soundness and Implementation. Florida Testing Profession-NEA and National Education Association, Washington, D.C., July 1978.

No one will deny the need for all students to read with a high degree of understanding and to compute with proficiency. But is it an injustice for teachers to assure students that the ability to read and compute as measured by contemporary tests, will lead to gainful employment?

If one of the purposes of schooling is to prepare students for the job market, are these skills sufficient, in today's economy. Are the unions willing and ready to accept new young people into their ranks? What preparations are realistic? How are these to be assessed?

Conclusion

Throughout this conference, reference was frequently made to the use of a teacher's professional judgement in determining student progress, achievement and placement in programs.

Is it possible, at this point in time, to train school personnel and establish credibility and dignity for alternative procedures so that public confidence is again placed in decisions made at the school level? - decisions based on sources of data generated by both pupils and practitioners and made on a longitudinal basis.

In summary, Sections I, II and III have attempted to describe specific examples of some overriding issues, processes, and concerns of testing in American schools. These include:

- How to provide the public with information about the nature of tests, access to tests, and reports of student achievement on tests.
- How to demonstrate to students the relationship between the world of work and school achievement, and establish the need for accomplishment in school.
- How to obtain a balance between changes in assessment procedures and the need to leave procedures in place long enough to inform the public, train staff, and orient students.
- How to change the negative reward systems in funding school programs - when students do well, funding is ended.
- How to reduce the high anxiety and negative social aspects of testing.
- How to relate pre-service and in-service education of school staffs to school assessment procedures.
- How to inform legislators, employers, and union representatives of the problems in using test data for personnel decisions.

IV. RESEARCH ISSUES

When considering the role of testing in educational decision-making it is necessary to be explicit about the type of decision to be made and the characteristics of the decision-maker. At the same time there appear to be several classes of problems that, at least nominally, cross out most of this complexity. The following represents an attempt to identify some of those cross-cutting classes of problems. First, however, two general areas of research offer potential for informing future research on testing and education decision-making.

A. Background Research

Much of the existing knowledge about testing practices in education is based on anecdotal evidence. A careful documentation and dissemination of the uses and abuses of testing would: (a) identify areas where efforts should be launched to bridge the gap between best practices and existing practices, and (b) suggest properties that should be built into tests of the future.

1. Systematic documentation of the uses and abuses of current testing practices, particularly "standardized tests."

This statement provides the context for the following five questions:

- a. How do current testing practices affect different kinds of students; e.g., ethnic groups, SES, cognitive style, social style, sex?
- b. How do current testing practices open/close opportunities for different kinds of students?
- c. What are the effects of mastery testing programs that keep students from moving on to new material?
- d. How do current testing practices impact on the curriculum?
- e. What is public reaction to test information?

2. Research on decision-making in education; e.g., by teachers, administrators, parents, the public.

Tests are but one source of information on which educational decisions are based. The utility of tests to aid decision-making should be viewed from this larger context. To analyze the adequacy of test information in comparison to other data, questions such as these need answers:

- a. What are the models which underly the several decision-making processes in education?

- b. What sources of data are used for what purposes; e.g., Lists A and B.
- c. What are the effects of adding a lottery component to decisions about allocation of resources?

B. Some General Problem Areas in Testing

1. Methods should be developed to make better and more complete use of information from sources other than formal testing. Tests should be evaluated in the context of what they can add to the information base that already exists.

2. Aggregation of Data

A test is a collection of small pieces of information about student achievement. If this information is to be used in decision-making it will be broken down in some way; e.g., across tasks for an individual or across individuals for a school. This suggests these research questions:

- a. What are the implications of different aggregation strategies?
- b. Who must be able to understand aggregated data?
- c. Should testing procedures be developed that allow different aggregation strategies for each user?

3. What kinds of safeguards are necessary to protect children when detailed student portfolios are maintained?

4. How can we improve information available to the public about student performance?

- a. What does the public want to know?
- b. How much information is useful?
- c. How should the information be presented?

5. How can the content validity of tests be improved?

- a. Achievement test development must include processes which ensure the subject matter integrity of the domain tested, by being concerned with both depth and difficulty.
- b. Can a single testing procedure serve equally well the several types of educational decisions?

c. How can the amount of testing be decreased without having serious effects upon the amount of useful information?

d. What are the effects of varying the format of test items? Should tests allow students to select the format on which they will be tested?

e. As a context for interpreting achievement test results, what procedures must be developed for reporting the curriculum to which students have been exposed?

6. What would a testing system look like which attempted to emphasize what each student can do?

a. Should students be given the option, at least in part, of deciding on what they will be tested?

b. How can tailored testing be used to increase the number of areas of achievement tested within the given period of time?

c. What are the effects on students of spending substantial amounts of testing time documenting in detail what a student has not yet learned?

V. CONCLUSION

The NIE should implement all of these research recommendations. It seems to us that a critical place to begin is with the support of research which:

1. describes students strengths and abilities using a variety of reasonably validated new assessment approaches, and

2. compares these descriptions with results obtained from standardized or norm referenced tests.

APPENDIX: THE NATIONAL CONFERENCE ON ACHIEVEMENT TESTING AND BASIC SKILLS

THE NATIONAL CONFERENCE ON ACHIEVEMENT TESTING AND BASIC SKILLS

March 1–3, 1978

U.S. Department of Health, Education, and Welfare
Joseph A. Califano, Jr., Secretary
Mary F. Berry, Assistant Secretary for Education

National Institute of Education
Patricia Albjerg Graham, Director

February 1979

"The spirited discussion of the times can help improve tests themselves by highlighting areas where further research is necessary, and can improve the use of tests by shaping our understanding of the appropriate and effective employment of achievement testing in the instructional process."

— *Jimmy Carter*
(from the President's message to the
National Conference on Achievement
Testing and Basic Skills
Washington, D.C., March 2, 1978).

FOREWORD

Defending the intrinsic worth of education is becoming a key challenge for our educational institutions. Our educational system is being called into question—not simply because there are fewer students to serve, but because the value of education is less clear than it once was.

To understand why the value of education is being questioned, we must look to our past. American society has tended to view our schools as a cure for every social and economic problem. When these problems have persisted, we have blamed the schools, not society.

We must recognize that the primary reason we educate people is not for salvation, morality, mobility, or even for economic gain. It is to make persons literate—able to read, write, manipulate symbols, and develop independent means of making judgments and determining actions. Until we are willing to argue that literacy, in its broadest sense, is adequate justification for the educational system, we will continue to charge that system with undertakings in which it cannot, by itself, succeed. And more importantly, we will continue to misunderstand the central purpose of education.

I am not taking the position that we should concentrate all our educational efforts on increasing literacy. We must also continue to work toward improved equity in education—toward reducing obstacles to academic achievement that result simply from a student's being nonwhite, female, or poor.

I believe these related concerns, fostering literacy and increasing equity, are the most important challenges for education today. Public policy toward education at the local, state, and Federal levels must come to view the ability to read, write, and think—to establish an independence based on knowledge of the world and self—as the preeminent reason for education and as a necessary condition of equity.

Patricia Albjerg Graham, Director
National Institute of Education

PREFACE

When HEW Secretary Joseph A. Califano, Jr., convened the National Conference on Achievement Tests and Basic Skills in March 1978, he focused the attention of professional educators and the lay public on an issue of national importance. Public concern about student achievement in the fundamental skills of reading, writing, and arithmetic has grown for more than a decade. During this period, scores on standardized achievement tests administered to college-bound students, as well as measures of student achievement in the basic skills, steadily declined. A number of individual researchers have offered competing analyses of the reasons for this decline. The College Entrance Examination Board, the National Academy of Education, and the National Academy of Sciences have each convened panels of experts in an effort not only to understand the nature of the test score decline, but also to examine how tests are used or misused and whether they accurately and adequately measure a student's educational progress.

To help develop a common understanding of the many dimensions of the testing issue and to provide a forum to discuss the Federal role in assisting state and local agencies, the March conference brought together teachers and educational administrators; experts on testing; government officials from the Federal, state, and local levels; parents; and community representatives. Harold Howe II, struck a tone of conciliation and shared purpose for the diverse group with his opening words, "This conference was called to talk about how the Federal Government's concern for schooling, centered both in the Congress and the HEW, can be joined to all the interests represented here to devise more useful ways of determining what is being achieved in schools."

Having identified the differing perspectives and opposing viewpoints among the participants, Howe appealed for a truce in the "four-cornered shouting match" on behalf of the indisputable need to improve the quality of education in America. The appeal carried, and the antagonism predicted for such a gathering was conspicuously muted. As testimony to the conferees' unity of purpose, small group discussions and corridor conversations found the conflicting interests coalescing in search of shared positions and practical solutions.

In this report of the National Conference on Achievement Testing and Basic Skills, we have endeavored to capture the sense of both the diversity and the consensus among the participants. Rather than providing a verbatim reproduction of the various papers and position statements, we have attempted to catch the underlying message of the proceedings by reporting salient extracts from the various presentations and summaries consistent with the speakers' points of view.

Most speakers agreed that, on the whole, the schools have done a creditable job of educating the American student. These same speakers also observed that the schools still must find ways to meet the special needs of groups that traditionally have attained lower levels of educational achievement. Most concurred, as well, that testing is an important tool in the evaluation of schooling, but that a better understanding of both the potential and the limits of tests is needed not only by the educators who use the tests, but also by the students who take them and the public who reads their results.

The recommendations developed during the conference emerged from the series of small group discussions that followed each plenary session. The recommendations touch on nearly every aspect of the educational system. In their completeness, they reflect a consensus of the participants about an important imperative that should be part of any activity in education: Remember the whole picture; remember the whole student.

The conference participants, both during the time of the meetings and in correspondence following them, made unusually significant and helpful contributions to the conference. The following proceedings reflect their sensitivity, sensibility, and judgment. We at NIE were pleased to organize this conference and, especially, to attract to it persons of such perception and commitment. We are in their debt.

Tommy M. Tomlinson
Maureen E. Treacy
National Institute of Education

CONTENTS

Page

FOREWORD

Patricia Albjerg Graham, Director
National Institute of Education v

PREFACE

Tommy M. Tomlinson and Maureen E. Treacy
National Institute of Education vii

PART I—ACHIEVEMENT TESTS AND BASIC SKILLS: The Issues and the Setting

Keynote Address by Harold Howe II, Vice President for Education and Research
The Ford Foundation 1

PART II—ACHIEVEMENT TESTING: An Overview of the Issues

Comments by Lauren B. Resnick, Codirector
Learning Research and Development Center
University of Pittsburgh 5

Vito Perrone, President
National Consortium on Testing 5

William W. Turnbull, President
Educational Testing Service 6

Bernard C. Watson, Vice President for Academic Administration
Temple University 7

Maria Ramirez, Assistant Commissioner
New York State Department of Education 8

Diane S. Ravitch, Assistant Professor
Teachers College, Columbia University 9

PART III—ACHIEVEMENT TESTING: The Interests

Comments by Ohio Lt. Gov. Richard Celeste 11

Albert Shanker, President
American Federation of Teachers
and
John Ryor, President
National Education Association 11

Ann Kahn, National Secretary
National Congress of Parents and Teachers 14

John W. Porter, Superintendent of Public Instruction
State of Michigan
and
Frederick H. Schultz, Chairman
Florida Education Council 14

Ramon C. Cortines, Superintendent
Pasadena, California, Unified School District 16

PART IV—BASIC SKILLS AND TESTING: The Congressional Perspective

Senator Claiborne Pell, Rhode Island 17

Senator S.I. Hayakawa, California 18

Congressman Michael Harrington, Massachusetts 19

PART V—ACHIEVEMENT TESTING AND BASIC SKILLS: The Federal Response

Joseph A. Califano, Jr., Secretary
Department of Health, Education, and Welfare 21

Patricia Albjerg Graham, Director
National Institute of Education 21

Congressman John Brademas, Indiana 22

Mary F. Berry, Assistant Secretary for Education
Department of Health, Education, and Welfare 22

PART VI—SENSE OF THE CONFERENCE: Conclusions and Recommendations

Martin Kaplan, Executive Assistant to the U.S. Commissioner of Education
Office of Education 25

APPENDIX: THE CONFERENCE AGENDA 31

I. ACHIEVEMENT TESTS AND BASIC SKILLS:
The Issues and the Setting

"Compared with the rest of the world, schools in the United States are doing a good job," said keynote speaker Harold Howe II, Ford Foundation vice president for education and research. "I do not know a country with a population as diverse as ours and so tumultuous an educational history that is doing as well as we are."

In setting forth the issues for the conferees, Howe acknowledged that American schools still have the task of meeting special needs of the handicapped, migrant children, the Spanish-speaking, blacks, American Indians, and students from other underachieving groups. But he commended "the very great achievement of American schools in bringing into the classrooms almost all the children of almost all the people.

"Our capacity to move on the two fronts of educational equity and excellence at the same time" he attributed to the diversity of institutions in the educational system and to the "freedom within a local school system to vary instruction and materials to fit its pupils." The essential means to improve quality in public schools is "helping teachers and principals to do a better job. . . . Changes in tests, textbooks, or promotion requirements are secondary lines of action that will not make much difference without the interest and support of teachers."

The Federal role even as a catalyst for change in the schools was resisted until the Elementary and Secondary Education Act of 1965, he said. This act recognized the diversity of both public and private schools and allowed schools to receive Federal funds without imposing uniformity. In contrast, he argued, "I find in present legislative proposals before Congress for national achievement testing an intent that opens the door to a nationalized school system." He criticized this concept as "anathema" and as a threat to "the diversity on which depends the balance between equity and excellence."

Howe noted that most states are active in some phase of achievement testing: Nineteen have adopted basic competency testing programs, sixteen are planning them, and eleven others are studying the matter. These programs involve two types of testing, Howe explained. One type will give "a composite picture at the state and local levels of the language and mathematics skills of students at varying ages in order to evaluate programs or to identify the need for remediation. . . . The other is designed to apply to a particular point in the educational progress of young people, either to qualify them for a diploma or promotion or to keep them where they are until they demonstrate more advanced skills."

He cautioned educators about possible effects of minimal competency testing: "The schools must become more willing than they are now to hold youngsters in grade and work flexibly and imaginatively to solve their learning problems, allowing them to progress through school at differing rates and with increased costs of school completion for a significant number of students. Or we must develop some alternative form of institution unlike the present schools and more like the Job Corps for those the schools are unable to serve. Or we must all live with the consequences of a group of young people rejected by our educational system and our economic system and embittered by their rejection in the name of improving educational standards."

He also raised several issues to be considered in the use of two common types of tests.

Norm-referenced tests, "which allow one to compare individual children and groups of children with each other on a common yardstick of achievement . . . take the temperature of the patient but give few hints about the disease or its diagnosis," he said. "One must know more than the test score to draw any meaning from the comparison it provides."

However, he added, as gate-keepers to determine whether a student should be promoted, their "convenience is vastly greater than (their) accuracy or fairness." Such tests are a legitimate part of promotion and admission decisions, he said, but they do not take into account other attributes of students, such as motivation and maturity, which also should be considered in decisions about students. He criticized schools and colleges for falling back on "the appealing neatness of a test score." Most agencies making and selling tests try to encourage their proper use, he said, but "unfortunately they don't succeed." He warned: "If parents really knew how inexact these beguiling numbers really are, they would be up in arms about their frequent misuse."

Criterion-referenced tests, he explained, allow one to determine whether individual children or groups of children have learned some particular knowledge or skill that is important to the school and measured by the test. But "one problem in their use is the difficulty of achieving agreement upon what students should be required to know or do."

He said it is important to continue improving criterion-referenced tests. "If we are to do a better job teaching basic skills to children, we need to improve the capacities of teachers to tune their teaching methods and materials to the needs of the individual child. To do this, we need very different instruments from the usual standardized achievement test; we need instruments that are carefully made, not hastily concocted from existing tests."

The effect on minorities of several aspects of testing should be carefully examined, according to Howe. "I have to raise the difficult question of whether the national mood for improving basic skills performance has hidden within it overtones of racism. There is no simple answer," he said. He agreed with those who have argued that "we do black children a disservice not to hold them to the same expectations we do white children. At the same time, we know that inequalities in the lives of children outside the school are reflected in school performance, and the black, Hispanic, and native American children are peculiarly subject to handicaps from both poverty and discrimination.

"Basic competency testing will certainly result in more educational delays and denials for minorities than for the rest of us. The only resolution of this difficulty I can accept is a clear commitment by those who advocate it to provide at the same time all the special additional services that may be needed to help less fortunate children meet the test requirements."

He suggested several possible Federal initiatives, including research on "the soft factors in judging students . . . those nonquantifiable aspects of human beings that must be considered in decisions about them."

Finally, he reminded the audience that tests are only part of the educational picture: "No amount of giving tests to children will ever teach children to read and write better."

II. ACHIEVEMENT TESTING:

An Overview of the Issues

"We have to consider where today's tests are failing and whether the functions of schools in which tests play a part can be better met, either with better tests, more informed use of existing tests, or different mechanisms altogether." With this charge, Lauren B. Resnick, codirector of the Learning Research and Development Center at the University of Pittsburgh, introduced a panel of speakers with diverse concerns about the uses of tests.

Resnick noted that standardized tests originally were adopted by school systems to increase the variety of school programs, to extend opportunities to students formerly denied them, and to assure acceptable standards of curriculum content. Now they are used for different purposes: to group students in a class or school before instruction begins, to "see how well we have succeeded in the instructional effort" at the end of schooling, and to monitor instruction on a daily or weekly basis.

These constructive purposes can have invidious consequences, however. Resnick explained that many have come to consider tests at least partly responsible for limiting the quality of instruction when school districts shape curriculum to match tests, for fostering inequality by creating rigid tracks that restrict learning possibilities, and for reinforcing through cultural biases the advantages conferred by birth and privilege.

Considerations in the Use of Test Results

Serious doubts that current testing practices are either very educational or informative were expressed by Vito Perrone, president of the National Consortium on Testing, a group that includes the Educational Testing Service, National Education Association, and other organizations and individuals concerned about testing.

Perrone reaffirmed the consortium's "vigorous support" of evaluation and quality education and strongly advocated a reappraisal of standards for both evaluating and using tests. "What we can measure competently with achievement tests, whether norm- or criterion-referenced, represents a very small subset of the goals of American education."

He suggested that the major weakness in proposals for accountability systems based on testing is that they generally ignore the people who teach and learn—those individuals whose standards are supposed to be upgraded. Further, "There is considerable ambiguity about the connection between achievement testing and basic skills," he said, arguing that success on achievement tests and knowledge of basic skills do not necessarily relate to one another.

"To suggest that more testing is a means for improving children's learning of basic skills is, at best, dubious. Connecting achievement testing and basic skills may well be good politics, but in the end, might make it difficult to bring the most productive focus to our deliberations."

He concluded, "In spite of what is claimed for testing today, much of it just hasn't been particularly useful to teachers, children, and parents—and this ought to be the bottom line."

A Test Publisher's Perspective

Accuracy, objectivity, and comparability were the three main virtues of standardized tests that led teachers—and later administrators and admissions officers—to adopt them widely when they were introduced, according to William W. Turnbull, president of the Educational Testing Service. "Teachers knew what an inexact business grading really is, and they welcomed the new development that held the promise of improving their information: The tests were consistently more accurate and more objective, and they showed how well the student did in comparison with pupils in the same or other classes or states.

"We had then the basis for a fine combination of techniques: standardized tests, which could measure sheer accomplishment in several areas very well, and teacher judgment, which could add dimensions inaccessible to standardized testing. . . . But we have since learned how difficult it is to keep achievement scores in that perspective—as one ingredient in a mix of information.

"Testing has undoubtedly suffered more from the excessive expectations of its most devoted advocates than from the attacks of its critics," he observed. He summarized three fallacies—"born of overenthusiasm"—that have led to misuse of tests. The "micrometer fallacy" attributes to test scores a precision—"an infallibility"—that they have never possessed. The "whole person fallacy" tends to cause much more to be read into achievement test scores than the amount a student has learned in a given subject. The "equal preparation fallacy" leads to beliefs that the test compensates for the differences in academic development of children whose learning opportunities have differed dramatically.

When these fallacies are resisted, he noted, tests do "a rather good job of measuring the academic accomplishments they purport to measure." However, they cannot measure the whole person nor compensate for differences in background. But "standardized test scores have given a new dimension of information to educators, primarily through comparability across geographic areas and across spans of time."

Turnbull identified three legitimate uses of standardized tests: "their use by teachers to determine how much the individual student has learned, by administrators to determine how much classes and larger groups have learned, and by university admissions people to discover how well prepared a prospective student may be.

"I suggest we put tests in a reasonable perspective and join forces to improve both the tests and their use," he concluded.

Implications of Tests for Children from Poor and Minority Families

It is the misuse of tests, not testing itself, that can be disadvantageous to students from poor and minority families, according to Bernard C. Watson, vice president for academic administration at Temple University. Watson made it clear he did not advocate abandonment of tests, but he expressed concern that tests not be expected to do things they simply cannot do. Misuse of tests "takes on special significance" for poor and minority children because, as groups, these children generally do less well on tests, he said.

"Tests can be helpful tools . . . but we know that any test or battery of tests measures a fairly narrow band of qualities, abilities, or other characteristics of individuals." But he estimated that "tens of thousands" of counselors, teachers, admissions officers, employers, and others think test scores really describe and summarize the essential and inherent complexity of the individual in question. Too many decision makers are poorly equipped to understand the limitations of test scores in evaluating human potential.

Part of the responsibility for this lack of understanding, he said, "must be placed squarely on the shoulders of test makers and the purveyors of tests." They have introduced caveats about what their products are designed to do and what they can and cannot do, he acknowledged; the information is there for the intelligent and interested to use. But it is neither highlighted nor read and understood by people who use test scores to make decisions that may dramatically affect an individual's life.

Watson also was critical of test makers for creating an impression that "a test can be developed to measure almost anything." He acknowledged efforts by professionals in the field to provide guidance for the users of tests, including the 1974 publication of *Standards for the Use of Psychological Tests* by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. "But how many counselors, teachers, or administrators have read or understood this document?" he asked.

Misuse can be blamed in part on a belief in the magic of numbers by a technologically advanced society, he suggested. "We believe that if we can take human qualities or characteristics, reduce them to quantitative terms, analyze them with sophisticated techniques, and come out with a correlation for a series of them . . . we've got it!. . . That's what we do with numbers like IQ scores, achievement test scores, grade point averages, and other things, but particularly with tests," he said. "It's a convenience, but it certainly does not represent reality."

Minority and poor families are particularly concerned about the use of test scores to legitimize such "labeling" practices as tracking, credentialing, and denying access. "My basic concern is what happens to individuals who become the victims of inappropriate uses of tests. . . . I'm concerned about anyone making decisions about any human being on the basis of something as limited as a single test score or series of test scores."

Finally, concern exists that test scores are not neutral—"they represent judgments about values important in this society," he said. From there, he continued, "it is not too great a leap in logic to suggest that test scores affect how much value interpreters of the scores place on the individual in question . . . in terms of potential growth, achievement, and access to the traditional symbols of success and status in our society."

Cultural Considerations in Achievement and Basic Skills

The effects of testing on children from ethnic and cultural minorities was of particular concern to Maria Ramirez, assistant commissioner for general education and curricular services, New York State Department of Education. There is no such thing as culture-free achievement testing, she said. The very act of achievement testing assumes a cultural value placed on achieving, but in the United States a number of subcultures do not value achievement quite as highly as the majority population. "Some groups are more cooperative than aggressive, more deferential than competitive. Rewards are primarily for the group, not for the individual. Thus, the first cultural bias of achievement testing is the very concept itself," she contended.

The question becomes not how to eliminate cultural bias from achievement tests, but how to be sure that the biases are valid. She had some suggestions:

- The area selected for achievement testing should be an important part of the culture in which a child has functioned or is to function. If the area is one not likely to have been experienced outside the school's curriculum, it should be an actual part of the curriculum. And if the area is part of the curriculum, it must actually be taught in the classroom of the student to be tested.

- Language knowledge and skills knowledge should not be confounded in the preparation and administration of tests. Tests of skills knowledge must be given in a student's first, i.e., most proficient, language while second language skills must be understood and measured for proficiency in the second language alone.

- Translation is not enough. Many monosyllabic English words become polysyllabic words in Spanish, and levels of difficulty may not be equivalent.

Origins of the New Testing Movement

The demand for a new emphasis on basic skills and the imposition of minimum competency requirements for high school graduation are the expression of the discontent that parents and public officials now feel toward the quality of education, said Diane Ravitch, an assistant professor at the Teachers College, Columbia University.

"The public's demand for better education and higher standards of performance should have been initiated by educators," she said. But far from leading the movement, educators have developed various strategies to evade the question or discredit those who raise it. She argued that educators are worried because they perceive that the real object of the movement is not so much to test the competence of the children as it is to test the competence of the schools. She suggested that educators recall some basics of education: "The purpose of education in a democratic society is really quite simple: to improve individuals and by doing so improve society. The purpose of tests is to help measure whether we're attaining our goal of improving the academic achievement of pupils."

Ravitch also summarized the findings of a report, "Improving Educational Achievement," prepared by a panel convened by the National Academy of Education at the request of the U.S. assistant secretary for education. The report concludes that:

- American life today requires mass literacy of a higher order than ever before; a greater number of people have achieved that higher level. However, as expectations of minimum literacy have risen, increasing numbers of people who have not learned to read fall farther and farther behind the standards.

- Standards of excellence have declined. These declines are attributed to a persistent de-emphasis of intellectual rigor in substance and methods.

- The spread of minimum competency testing is attributed to a lack of trust felt by parents and legislators in the judgment of those operating the schools; but minimum competencies might well become accepted maximum competencies, resulting in a further erosion of educational standards.

- The consequences of automatic promotion must be re-examined. Learning problems are likely to go undetected, so large numbers of children move on to high school where their inadequate preparation leads to discipline problems and hampers the regular work of the school.
- Tests should be used to identify children who are not learning, to diagnose their special learning difficulties, and to direct additional resources and individualized attention to them as early and as often as needed.

III. ACHIEVEMENT TESTING:
The Interests

While representatives of national teacher and parent organizations stressed the need to improve testing methods, state and local educators recounted their experiences in successfully implementing assessment programs. As Ohio Lt. Gov. Richard Celeste noted in his introductory remarks, these educators brought "a very practical perspective, a contrast of expectations and experiences, and a glimpse at what really happens when we begin to look at achievement at local and state levels."

Teachers and Achievement Testing

The presidents of the American Federation of Teachers (AFT) and the National Education Association (NEA) expressed different views on the question of standardized tests. The NEA has called for a moratorium on them; the AFT believes they should be retained, and that the public should be educated as to what the test scores mean—and don't mean. Both presidents called for improvements in testing, and both defended public schooling from much of the criticism they see as implicit in the move toward minimum competency tests.

AFT President Albert Shanker observed that there is no universal "teacher view" on testing, that one would find "the same range of disagreements in a teacher organization as there is within a community or on a school board." But Shanker also stated that it would be unreasonable for a spokesperson for any such organization to recommend that testing be abandoned altogether.

Shanker's position is that the public wants to know how well students and schools are performing. If teachers totally object to testing as a means of measuring performance, parents will wonder what the teachers are trying to hide, and taxpayers may refuse to support educational programs.

However, Shanker underlined the anxieties about standardized testing often expressed by teachers. Tests take time away from teaching and have little effect on actual learning. Test results are frequently misinterpreted by the press and the public. The misinterpreted results may be used as "a form of simplistic accountability" in which teachers' jobs and salaries depend on their students' test scores.

Shanker suggested that the testing controversy is symptomatic of a greater issue: the loss of faith in public education. He attributed this problem to several factors:

- Back when few Americans were well educated, teachers and the educational process were held in high esteem. Now that public education

414

has succeeded in reaching the general population, teachers are seen as equals. Many people no longer look up to teachers, but rather look down at them.

- Counter-culture criticism denouncing public education has significantly affected public opinion during the last two decades. Yet the counter-culture does not seek to reform the schools, but rather to rebuke the values of society as represented by our educational system.
- The educational establishment has been too defensive. While condoning all the attacks on standardized testing, it has conveyed the impression that it doesn't know how to measure progress or make rational determinations. Then the taxpayers revolt because the schools don't demonstrate their effectiveness.
- School systems have succumbed to the pressures to encourage "innovative" programs while neglecting to support established methods.
- The new "open" policies which require test scores to be released to individuals and the public have not been accompanied by educational policies to explain the meaning of those scores.

To counter these problems, Shanker recommended a reaffirmation of the educational profession. "We have to start with the notion that it is possible to develop a body of knowledge, to develop a model of competent teaching practice over a period of time." He characterized education as an artistic profession with a base that is part science, part experience, and part experiment. Establishing goals of student achievement and preparation for life in a competitive, achievement-oriented society would help to attract similarly inclined teachers to our system, Shanker concluded.

On the testing of teachers, Shanker noted: "I don't think there is any test that will tell us whether a person is going to be a good teacher—that we will find out later—but tests can tell us whether a math teacher knows enough math and an English teacher enough English."

In summary, Shanker recommended the improvement rather than the rejection of standardized tests: "You can't get rid of the unemployment problem by firing the Bureau of Labor Statistics." He suggested a federally sponsored public and professional information program for "truth in testing" to counteract the problems of misinterpretation. Finally, Shanker defended public education: "It's hard to say that we're illiterate and failing as a country while everybody else in the world is working to provide us with goods in exchange for our know-how. If things were as bad as some say they are, we wouldn't be where we are now."

NEA President John Ryor emphasized the confusion and contradictory attitudes held about schooling: "Parents who almost universally reject the notion of a national curriculum . . . at the same time embrace the need for a national standardized test—without understanding the relationship between the two." Perhaps we should answer a long series of questions before we start devising further means of evaluating, he said. "In my view, standardized evaluations of education in the United States make no more sense than insisting that education in Point Barrow, Alaska, ought to be identical to education in White Plains, New York."

Since 1971, the NEA has sought a moratorium on standardized testing because of beliefs that the tests do not do what they purport to do, that they tend to be culturally biased, that they are norm-referenced, and that they automatically label half the students as losers. Standardized tests seldom correspond significantly to local learning objectives, and they can't be used to measure growth over a short period of time, Ryor argued. Further, many school systems tend to misuse the tests in jumping to unwarranted conclusions about curriculum and to justify tracking students into inflexible decisions regarding education and career. Countering the common accusation that such reasons for teachers' opposition to standardized tests are self-serving, Ryor contended that "maintaining an evaluation model based on nationally normed tests would be the easiest of all possible worlds for teachers, as there's great comfort in anonymity and that's precisely what standardized tests provide."

After three years of study, the NEA Task Force on Testing concluded that tests should be used primarily to improve instruction, to diagnose learning difficulties, and to plan educational activities in response to those learning needs, but not in ways that deny any student full access to educational opportunity.

The essence of education "lies in what happens between children and their parents and teachers and classmates," Ryor observed. "These relationships are delicate and susceptible to strong outside influences. An accountability system must take care not to damage these relationships and, above all else, must be acceptable to those who are expected to abide by it, for in a pluralistic society an accountability system must promote diversity, not conformity."

Ryor reaffirmed NEA's belief in the importance of evaluation and supported such tools as individual diagnostic tests, teacher-made tests, school letter grades, and criterion-referenced tests—but not standardized tests.

"The whole notion of norm-referenced, standardized tests makes a lie out of the often-stated concern for individual differences. . . . The most tragic aspect of such tests is that they contribute to the training of children to do better than somebody else. In my view, the only competition worth the name is competition with oneself," Ryor said.

"A free society needs, above all things, a free and learned citizenry, and the first task of education is to stimulate curiosity, to teach children how to learn and how to remain open. If we can't do that, then we can't truly educate; we can only train. And the difference between training and educating is monumental. Educated people remain curious for a lifetime, but one who is trained only performs . . . a nation of trained people could not possibly be learned and flexible enough to meet all the complex challenges that lie ahead of us."

A Parent's View of Achievement Testing

Ryor's concerns were echoed by Ann Kahn, Fairfax County, Virginia, school board member and national secretary of the National Congress of Parents and Teachers. "I come to you reflecting a deep and growing concern about the use of norm-referenced achievement tests and the virtual con job that I think school systems have done on most parents regarding the meanings of those test scores.

I'm appalled at the dependence of schools and parents on a rather meaningless accumulation of data. I know our dependence on years of norm-referenced achievement tests has simply obscured the information that we need to know; it has delayed the responsibility that we need to assume to be really accountable for the teaching of children."

Kahn went on to point out that the comparative data resulting from norm-referenced tests do nothing to help determine what individual children know and don't know, provide no precise information to schools about whether children are learning what schools are attempting to teach, and suggest no basis upon which parents and school board members can make judgments about the effectiveness of programs or the quality of teaching. Where comparative data are needed, she recommended use of limited standardized achievement testing with valid national sampling techniques. Kahn also recommended using locally formulated criterion-referenced tests, educating principals and teachers to improve use and interpretation of measurement, and, helping responsible school districts find alternative ways of getting information. In addition, she encouraged Federal and state governments to use other than test criteria to determine eligibility for funds for special programs. Finally, she urged, "Go back to the community and unsell the job that has been done over the years by educating the parents about the limitations of norm-referenced testing."

Two State Assessment Programs

State minimal competency programs have worked well in two states, according to panelists John W. Porter, superintendent of public instruction for the State of Michigan, and Frederick H. Schultz, chairman of the Florida Education Council.

"We have a lot of children who for various reasons are not able to be competitive in the American way of life. What the public is asking us is whether or not there is a way within the American public school system to enable

417

these youngsters to be competitive, at least at a minimal level," Porter said.

Michigan responded to that public concern with a statewide education assessment program, developed over the past ten years. "The basic purpose of the assessment program is to measure educational needs against expected learning. The teachers write the test items and have told us what they think are reasonable objectives for all students in the state," Porter explained.

Minimum performance objectives have been developed in ten distinct areas such as career development, personal growth, and social understanding, as well as in reading and math. As a matter of public policy, the state assumes that, once the needs of youngsters have been identified, those special needs can be met. Proposals for experimental and demonstration programs are required of systems where the test data demonstrate special needs; compensatory funds are made available.

According to Porter, the old system of standardized testing based on a normal distribution curve was good for its time, but was based on four assumptions that should be considered invalid now. "For 200 years it was assumed that the operation of our public schools would be relatively inexpensive; that our public schools would be able to screen, sort out, and select; that those who owned real property and paid for the schools would also be the parents of the children; and that the 'have-nots' would not actively participate. Those assumptions simply are no longer feasible."

Additional advantages of the Michigan program outlined by Porter include the built-in necessity of studying expectations in advance of instruction, a capacity to report to parents the performance of their children on an item-by-item basis, its encouragement of professional staff at the classroom level, its automatic check on the purchase and use of materials, and its potential to focus attention on the specific needs for in-service programs.

Speaking on the Florida program, Schultz highlighted the events leading to passage of Florida's minimum competency requirements—among the first in the Nation to be implemented. The Florida Educational Accountability Act of 1976 mandates the administration of basic skills competency tests in grades three, five, eight, and eleven, and the attainment of a minimum score on a special fundamental literacy test to qualify for a high school diploma.

Of the approximately 100,000 students who took the functional literacy examination for the first time in the fall of 1977, 36 percent failed the computations sections and 8 percent failed in the communications section; 37 percent of the students failed one of the two sections of the examination.

The reaction to the results? "A roaring success," according to Schultz. "If we put the test program on the ballot, I believe it would receive at least 80 percent favorable vote; as far as I know there is no major officeholder who has opposed it. As far as I know, there is not even a major candidate for public office who has opposed it, remarkable as that may seem."

Results of the test, he added, re-emphasized the need for remediation, increased the activity of schools that reported fairly high percentages of failure, and stimulated tremendously the degree of parent involvement in education.

Perspectives from a District Superintendent

Standardized test scores had been declining in Pasadena for about five years, but the trend had gone virtually unnoticed until a 1970 court order to desegregate, according to Ramon C. Cortines, superintendent of the Pasadena, California, Unified School District.

At that time the national news media became highly interested in knowing how the students were doing. After reviewing probable causes for the test scores, the board of education voted unanimously for a resolution stating that the first priority of the school district would be to reverse the decline of SAT scores of all students.

"As I look back, there were some particularly noteworthy points in relation to improving test scores for minority children, for poor children, in fact, for all children," he recalled.

An ongoing evaluation process was implemented for sequential math and reading programs. Consistency in the amount of time spent on reading was emphasized for students scoring below grade level. In-service training was developed to improve teachers' ability to interpret student needs. Interruptions of reading and math instruction were eliminated. Title I programs were incorporated into the core curriculum instead of being treated as a program appendage. Staff members were encouraged to expect all children to learn, and parent involvement was encouraged. Auxiliary services were introduced in health counseling, psychological evaluation, library science, and other areas.

The Pasadena Board of Education now reviews achievement data annually and holds community meetings where principals review the meaning of the data for parents and staff. The district has initiated a Saturday reading and math program using a certified teacher, assisted by a competent aide and selected high school students. A similar program is offered after school in bilingual instruction and English as a second language. In addition, in-service training was instituted to familiarize administrators with the math and reading sequences.

Cortines attributes a large part of the success of the effort to "being honest with the staff and the community and opening a two-way dialogue." He concluded, "I am happy to report that a multiracial, socioeconomically diverse school district can say that its standardized test scores are not on the decline, but improving. In some grades and in some areas they are doing so very markedly. Even in primary grades there are some signs that the achievement gap is being narrowed between majority and minority children."

419

IV. BASIC SKILLS AND TESTING:
The Congressional Perspective

As the public has become more concerned over testing in the schools and over the declining math and reading skills of students, more attention has been paid to these issues in Congress. Senators Claiborne Pell, Rhode Island, and S. I. Hayakawa, California, and Representative Michael Harrington, Massachusetts, offered their views to the conference on legislative aspects of testing and basic skills.

Comments of Senator Claiborne Pell

Senator Pell expressed concern over the "spotty performance" of youngsters in writing and math in the National Assessment of Educational Progress, as well as the continuing decline in SAT scores by prospective college students. He agreed with the findings of the National Academy of Education panel on testing and basic skills that a number of sociological influences have affected the decline. "Obviously the society at large, not just the schools, plays a major role in achievement decline," he said.

Senator Pell discussed a proposal for a Council on Quality Education as a means by which the Federal Government might help states and localities combat the decline. Originally proposed by Adm. Hyman Rickover in testimony before Pell's Subcommittee on Education, Arts, and the Humanities, the council would be a blue-ribbon group commissioned to develop Federal standards for performance at various grade levels. The proposal also calls for the development of a national achievement test that would measure student performance against a national standard and could be voluntarily adopted by local and state school boards.

Pell told the assembly that, in his judgment, every program of testing that emerges from his subcommittee this year would contain neither provisions for the use of test data as a basis for the distribution of Federal funds nor mandatory requirements for testing or test development. The point of such a program would rest solely on the voluntary development and implementation of methods to assess educational performance.

Sensitive to the resistance parents and educators often express to a Federal presence in the classroom, Pell acknowledged that the undertaking would "have to be carefully meshed with ongoing state efforts." Following these comments, the Senator asked the conferees whether they would favor such a test. In a dramatic affirmation of the anticipated resistance, they were nearly unanimous in their opposition to his proposal.

Pell also suggested that Federal assistance could be given to states and localities to develop their own tests, and encouraged the NIE to work toward that end. A diversity of tests within a state could reduce the possibility that teachers might "teach to the national test," he noted.

"We must not lose sight of an important point," Pell said. "Tests and test results themselves do not improve student performance. . . . All tests do is tell us the extent to which schools have been successful in teaching specific students specific skills. . . . What is important is what teachers and parents do with the test results when they have them."

Comments of Senator S. I. Hayakawa

Senator Hayakawa reported that, "during the last year, Americans spent far more on all levels of education than they spent for defense—for primary and secondary education alone the bill was $75 billion, more than four times the expenditure of 1960." But judging from his mail, he noted, parents seem to feel the more they pay the less they get. Thus parents have turned to the Federal level for help.

The Senator proposed several possible Federal actions:

- Pursue a method to identify and help early in their schooling students who have difficulty with basic skills. He specifically suggested developing a test of competency against clearly defined standards for reading, writing, and mathematics at intervals throughout schooling to insure that the student is acquiring these necessary skills. Such a test "would be made available to those states and localities requesting it to be used as they see fit," he said.
- Encourage setting high rather than low standards to avoid institutionalized incompetency for future graduating classes.
- Encourage teachers to assist in developing competency standards. He said teachers should welcome the opportunity to be held accountable, because the alternative is likely to be a tremendous increase in the number of malpractice suits being filed.
- Share the responsibility for conducting remedial programs to bring students up to specified standards.

Hayakawa also responded to concerns expressed earlier in the conference about cultural bias. "I'm not so sure that cultural bias is all that serious a matter," he said. "All testing inevitably has to have a cultural bias. If you're being tested in China, you're going to be tested in Chinese, whether you like it or not, and so in American culture I think we shouldn't hesitate to have a cultural bias that says 'This is America.' We are one nation, not fifty nations, and we desperately need a nationally calibrated thermometer of educational achievement."

Comments of Congressman Michael Harrington

While expressing strong feelings about the need for improving education, Representative Harrington expressed reservations about reliance on testing as a quick-fix solution. "While I do not partake of the hysteria which in some cases would have us believe that most students now graduating from high schools are functionally illiterate," he said he was "also hard put to be as optimistic as Harold Howe," the conference keynote speaker. "I need only return to the schools in my district, meet with the students themselves, and talk with concerned parents to realize that there are many out there who are being failed by the educational process and that something needs to be done about it."

However, Harrington added, "I also believe that it is a terrible mistake to waste crucial time and resources turning to standardized tests for the answers. The real answers, if they can be found, lie in difficult assessments of social, economic, attitudinal, and educational components that make our schools what they are today. These components include inadequate funding, which President Carter's new proposal for increased Title I money for impact assistance may help to ameliorate. They include, as well, the nature of discrimination in society, employment, and housing, which has cut off both minorities and whites in the lower income brackets from American life as we teach it in the schools. They include the climate in the schools, which are often the grim centerpieces in miles of urban blight—schools where crime, discontent, and alienation often prevent a teacher from focusing on the subject. They include deteriorating relationships between indifferent students and disillusioned or ineffective teachers. These are the real issues and our grappling with them productively is only forestalled, if not completely circumvented, by our preoccupation with test scores," he said.

"The public is presented with score declines and sees them as a confirmation of fears about the quality of education. So, the clamor for more tests to maintain standards. The result is that legislators and school boards, quick to respond to public complaints with highly visible remedies, begin to mandate tests of all sorts to expand our evidence of the problem and enforce minimum requirements. I would be innately suspicious of any new trend which thirty-five states rush to join within three years." He added a final caution: "There is the clear danger that minimum tests will become maximum tests."

In an effort to create a better public understanding of testing and how tests should be used, Harrington proposed that an independent testing commission on the Federal level be given access to all the relevant data from standardized tests and charged with disseminating detailed information—in layman's language—pertaining to predictive and content validities, reliability, socioeconomic bias, and other factors for the benefit of parents, educators, and policy makers alike.

V. ACHIEVEMENT TESTING AND BASIC SKILLS:
The Federal Response

President Carter's message to the conferees stressed the Administration's concern:

> There is no greater challenge than the one facing our educational system—to ensure that all children learn, at the very least, to read, to write, and to compute.

Reaffirming and expanding on this message to present the Federal response to issues regarding achievement tests and basic skills were Joseph A. Califano, Jr., secretary of the Department of Health, Education, and Welfare; Patricia Albjerg Graham, director of the National Institute of Education; Congressman John Brademas, Indiana; and Mary F. Berry, assistant secretary for education, Department of Health, Education, and Welfare.

Comments of Joseph A. Califano, Jr.

In conveying the President's message to the conference, HEW Secretary Joseph A. Califano, Jr., added his own interest in improving the Nation's educational system. He announced one Federal initiative already begun—the establishment of a new office of testing and assessment within the National Institute of Education.

While the new office will not pretend to have all the answers, he said, "Its principal role will be to help people with questions find places with answers throughout the nation: education organizations with experience in testing, universities and scholars who are experts on tests, and states or school districts which have adopted successful testing programs."

Comments of Patricia Albjerg Graham

The Administration's position on and commitment to education was emphasized by NIE Director Patricia Albjerg Graham. Addressing some of the more specific questions raised at the conference, Graham said, "Testing is not a central issue to the National Institute of Education. Increasing student achievement is the central issue, and testing is simply a means by which we can gauge whether we are achieving that end."

She repeated her opposition to a national test of academic achievement and noted that activities that are reasonable for the Federal Government to undertake include "research, technical assistance, and increasing understanding about testing."

Comments of Congressman John Brademas

Congressman John Brademas, chairman of the House Subcommittee on Select Education, emphasized that: "For the first time in eight years, there is a commitment to support education at both ends of Pennsylvania Avenue." The President and Congress are "talking now not as warriors but as negotiators, trying to figure out how much and for what purpose," he said.

Highlights of the pending legislative proposals for Federal activity in education include:

- An additional $400 million for the Title I program and, with a matching fund formula, encouragement to the states to implement their own compensatory education.
- A basic skills and quality education act, aimed at helping states develop basic skills demonstration projects, with competency and achievement testing, and programs to involve parents more directly in the education of their children.
- New initiatives in bilingual education, assistance to school districts that are desegregating, changes in the impacted aid program, and strengthening the use of private schools.
- An increase of $900 million (15 percent) in Federal spending—the largest increase in a decade—for elementary and secondary education.
- The Middle Income Student Assistance Act, which builds on existing programs to provide more higher education grant and loan opportunities for students of middle income families.
- A commitment to seek a cabinet-level department of education.

Comments of Mary F. Berry

Assistant Secretary for Education Mary F. Berry reviewed Federal actions that had already been taken to address problems of testing and basic skills. Namely:

- In the fall of 1977, a task force to focus on the improvement of basic skills was established within the Department of Health, Education, and Welfare; the task force includes representatives from all education divisions.
- During the past year, NIE sponsored a series of four regional conferences to discuss minimum competency testing.

- The National Academy of Education was commissioned by the assistant secretary for education to review policy questions germane to achievement testing and to make recommendations for appropriate Federal action.
- Support has been provided for the National Academy of Sciences to prepare a comprehensive analysis of testing theory and practice throughout the United States.

Summing up the Federal concerns, Berry said: "This is a crucial time for discussing the issues of testing. Professionals are divided over the use of tests; while some support the extensive and increased use of tests, others believe that many tests measure the wrong things and are culturally biased, and that results of tests are frequently misleading and misused. But whatever controversies exist among the professionals, public support for testing programs has probably never been stronger in this country.

"Whether or not test results are truly accurate measures of achievement, they are widely publicized and discussed because they are accepted as being important and uniform measures of how schools and students are performing. The question before this conference is not whether there ought to be tests, but what kind of tests there ought to be and to what uses they should be put. Our goal is to assure that tests are developed in such a way that they tell us something meaningful about individuals, that they promote quality in education, and that they help people to learn. We want to help educators at the state and local levels to use tests constructively and effectively in the education process."

VI. SENSE OF THE CONFERENCE: Conclusions and Recommendations

The lively small group discussions that followed the plenary sessions reflected the participants' keen interest and concern over the issues. In summarizing reports from all of the groups at the end of the conference, Martin Kaplan, executive assistant to the U.S. commissioner of education, offered several overarching consensus statements:

- The sense of this body is that there should not be, under any circumstances, a national achievement test.
- The Federal Government and the Nation should not lose sight of the big picture. Testing is only one part of evaluation, testing and evaluation are only part of schooling, schooling is only part of learning, and learning is only a part of our national effort to improve the quality of life.
- Schooling is everybody's business—parents, students, teachers, administrators, school board members, legislators. It is important to involve everybody at all stages in the process of decision making.

Regarding the general climate of education today, the emergent sense of participants was that:

- The schools are not disasters. This country has reason to be proud of its schools.
- Equal access to education has not been accomplished; it must be aggressively pursued.
- Our school system possesses a number of fine characteristics of our Nation; these characteristics include autonomy, diversity, pluralism, local control, and regionalism.

Specifically regarding testing, the small group rapporteurs reported the following expressions of positions and opinions:

- Testing is part of a larger consumer and accountability movement. One cannot talk about testing without talking also about the need for education to be accountable for achieving its goals.
- Testing and evaluation are useful tools for some purposes. They should not be eliminated merely because they are the messenger of bad tidings.

- The back-to-basics movement often has the tone of discrimination and reduced expectations for minority students. Tests must not be used unfairly for sorting, tracking, enforcing credentialism, labeling, or role typing.

- Education goals should drive the agenda of testing and evaluation, and not the other way around. Tests should not drive the curriculum. They should not force a separation of skills and content. They should not abet anti-intellectualism.

- We must remember what testing is not. More testing does not necessarily mean a better understanding of the whole person, or that more is learned, or that students are more accurately evaluated in terms of their own particular backgrounds, or that the basic skills have been better mastered. Testing does not necessarily mean anything about the ends of education. Testing is a means of assessing the acquisition of a valued end; it is not the end in itself.

- Instead of asking how we can make better tests, we should ask how we can make better use of tests, how tests can be effective means of diagnosis, and how that diagnosis can in turn be translated into educational intervention, remediation, and compensatory efforts.

The following compilation of recommendations for Federal activities includes those most often suggested by the groups. They are in four general categories: policy, implementation, legislative/executive, and research.

Policy

The Federal Government should lead the effort to create an atmosphere of open discussion and mutual understanding for state and local officials, teachers, and parents. These groups should continue developing, implementing, and interpreting testing efforts of their own choosing. Toward that general end, the conference participants recommended the following specific initiatives:

- Encourage awareness of the large view of education in our society—a view in which testing is only a small part.

- Promote appropriate expectations about the uses and limitations of tests and testing.

- Encourage involvement of the broadest possible constituencies in discussions of issues, formulation of standards, and development of solutions to problems.

- Sustain the Federal effort in compensatory education; eliminate the kinds of erratic support that cause programs to move quickly into and out of existence.

- Support and emphasize programs that lead to early detection and correction of learning difficulties.
- Resist all temptations to tell state and local agencies what to do about tests, about standards, about criteria, about techniques, and about the content of learning.
- Award funds and allocate resources based on data other than standardized test results.
- Keep reports of conferences, formulations of policy—in fact, all communications—moving in complementary and consistent directions through the information channels connecting local, state, and national levels of activity.
- Do not smuggle in tests. Do not do it through Title I. Do not do it through the back doors of other Federal agencies and programs.
- Put aside all ideas on behalf of a national curriculum or a national test of achievement.

Implementation

The Federal Government should take immediate action to bring about the following:

- More coherent evaluation mechanisms for federally sponsored programs.
- Models of comprehensive evaluation systems for use at the state level that provide an integrated picture of efforts from testing and diagnosis through remediation.
- The identification of schools that have been successful in contending with low achievement, and wide dissemination and replication of the means such schools employed.
- More research money to local school systems, state education departments, and tribal and territorial governments for the development of appropriate tests.
- A layman's summary of this conference and all other related forums for use especially by those who buy tests and use test results.
- More conferences, training workshops, forums, and publications for parents, teachers, journalists, administrators, local school board members, state legislators, counselors, community members, students, and experts.

Legislative and Executive

The Federal Government should exert a concerted effort to consolidate and strengthen the work already being done in the area of testing by a large number of offices, programs, agencies, and organizations. Suggestions include:

- That Congress re-enact the Education Professions Development Act, which elapsed in the mid-1970s, to provide re-education in testing for administrators, principals, counselors, and school board members.
- That Congress charter the National Academy of Education to serve as adviser to all interested parties in matters of testing.
- That Congress establish a special five-year, blue-ribbon commission to develop Federal standards for performance at specified grade levels and a council on quality education to develop a voluntary high school proficiency examination and Federal certificates of achievement.
- That Congress enact a basic skills and quality education act to assist states with the development of demonstration projects.
- That all existing education programs at the Federal level be consolidated within a single department of education.
- That a special office of testing and measurement be established within the National Institute of Education.

Research

Conferees urged the Federal Government to expand the scope of its research effort to seek more accurate answers to questions such as:

- How do people learn?
- What changes take place in learning over the course of a lifetime?
- What factors are involved in causing some people to succeed and others to fail at learning?
- What skills are basic in American society?
- How do all the different parts fit together and relate to one another in a learning environment?
- What do specific kinds of tests actually test?
- What are the appropriate uses for each of the different kinds of tests?
- How can test data be made more comparable within and across schools, districts, and states?

- What can be done to provide better evaluation of tests presently in use?
- How can the reporting of test results be improved?
- What information should be included in users manuals in order to make optimal use of associated tests?
- What can be done to make tests fair for particular minority populations within American society?
- How can tests be used to provide better diagnostic assistance for individuals and groups of individuals?
- To what extent do tests determine the selection of curriculum, and to what extent does curriculum determine the selection of tests?
- What educational practices follow as a result of the installation of minimal competency requirements and the reporting of minimal competency testing results?
- What do we know about the validity and effects of teacher-constructed tests?
- What is the effect of adult educational politics on pupil educational performance?
- What are the alternatives to testing for evaluation, diagnosis, and prediction?

APPENDIX: THE CONFERENCE AGENDA

Loews L'Enfant Plaza Hotel
Washington, D.C.

March 1–3, 1978

March 1, 1978

8:30–9:00 p.m.

Mary F. Berry
Assistant Secretary for Education

Overview of the Federal Program

9:00 p.m.

Harold Howe II
Vice President for Education and Research
The Ford Foundation

Tests and Schooling

March 2, 1978

8:30–10:00 a.m.

Panel I: ACHIEVEMENT TESTING: WHY AND WHEREFORE

Moderator:

Lauren B. Resnick, Codirector
Learning Research and Development Center
University of Pittsburgh

Panelists:

Vito Perrone, President
National Consortium on Testing

Considerations in use of test results

William W. Turnbull, President
Educational Testing Service

Standardized tests: Current uses and limitations

Bernard C. Watson, Vice President for Academic Administration
Temple University

Views on the use and misuse of tests: Implications for children from poor and minority families

Maria Ramirez, Assistant Commissioner for General Education and Curricular Services
New York Department of Education

Cultural considerations in achievement testing and the basic skills

Diane S. Ravitch, Assistant Professor of History and Education
Teachers College
Columbia University

Tests in a social functions perspective: What we have and what we need

10:00–10:30 a.m. Commentary from the floor

10:30–10:40 a.m. Charge to Small Groups

11:00–Noon Small Group Discussions

Topic: Achievement Tests and Basic Skills: Preliminary Considerations of the Federal Role

12:15–1:00 p.m. Remarks by *Secretary Joseph A. Califano, Jr.*

1:00–2:00 p.m. Teachers and Achievement Testing

John Ryor, President
National Education Association

Albert Shanker, President
American Federation of Teachers

2:00–2:45 p.m. Commentary from the floor

3:15–4:15 p.m. Panel II: ACHIEVEMENT TESTS AND EDUCATIONAL QUALITY: STATE AND LOCAL PERSPECTIVES

Moderator: *Lt. Gov. Richard Celeste*, Ohio

Panelists: *John W. Porter*, Superintendent of Public Instruction, Michigan

Test scores and educational policy

Frederick H. Schultz, Chairman
Florida Education Council

State legislatures and testing programs

Ann Kahn, Fairfax County School Board and National Secretary, National Congress of Parents and Teachers

A parent's view of achievement testing

Ramon C. Cortines, Superintendent of Schools
Pasadena Unified School District

Perspective from the district superintendent level

4:15–4:45 p.m. Commentary from the floor

7:00–8:00 p.m. Remarks from *The Honorable Claiborne Pell*,
Rhode Island
United States Senate

8:00–9:00 p.m. **Panel III: ACHIEVEMENT TESTS AND EDUCATIONAL QUALITY: A NATIONAL PERSPECTIVE**

Moderator: *Willard Wirtz*, Chairman of the Board
National Manpower Institute

Panelists: *The Honorable S. I. Hayakawa*,
California
United States Senate

The Honorable Michael Harrington,
Massachusetts
U.S. House of Representatives

March 3, 1978

8:00–9:00 a.m. Remarks by *The Honorable John Brademas*, Indiana
U.S. House of Representatives

9:00–9:10 a.m. Charge to Small Groups

9:15–11:00 a.m. Small Group Discussions

Topic: Achievement Tests and Basic Skills: Guidelines for Federal Assistance

1:15–1:45 p.m. Sense of the Conference: Overview of the Conclusions and Recommendations of the Discussion Groups

1:45–2:30 p.m. Commentary from the floor and Federal response

Chairperson: *Mary F. Berry*, Assistant Secretary for Education

Ernest L. Boyer, U.S. Commissioner of Education

Patricia Albjerg Graham, Director
National Institute of Education

*U.S. GOVERNMENT PRINTING OFFICE : 1979 0-631-562/2900